ISSN: 1692-7257 - Volume 2 – Issue 48 - 2026

ORIGINAL

RCTA
Revista Colombiana de Tecnologías de Avanzada
UNIPAMPLONA

A cascading hybrid model for assigning reviewers to scientific articles

Modelo hibrido en cascada para la asignación de revisores en artículos científicos

Liseth Paola Claro Ascanio[1]
Cristian Camilo Tirado Cifuentes[1]

[1] **Universidad de La Salle, Facultad de Ingenierías, Maestría en Inteligencia Artificial en Profundización, Bogotá, Distrito Capital, Colombia.**

Received: 11 May 2026. Revised: 27 June 2026. Accepted: 09 July 2026. Published: 16 July 2026.
Corresponding Author: Liseth Paola Claro Ascanio

How to cite: L. P. Claro Ascanio and C. C. Tirado Cifuentes, "Modelo hibrido en cascada para la asignación de revisores en artículos científicos", *Revista Colombiana de Tecnologías de Avanzada (RCTA)*, vol. 2, no. 48, jul. 2026, art. no. 4492, doi: 10.24054/rcta.v2i48.4492

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Abstract:

Scientific journals face delays in publishing new issues due to a shortage of expert reviewers and an increase in article submissions. This study proposes a combination of models with a cascading desing that utilizes natural language processing (NLP) and machine learning (ML) tecniques to automate the assignment of reviewers. A quantitative approach with an experimental design was adopted, using historical records extracted from the OJS platform of a real scientific journal in the field of Social Sciences, focusing on the disciplines of Business administration, accounting, economics, pedagogy applied to business development, entrepreneurship, local development, innovation and technology (2021-2025). The model is structured in three sequential stages: abstract acceptance classification, disciplinary field classification, and reviewer speed classification. The results were as follows: 100 combinations of textual representations, classifiers, and data balancing techniques were evaluated using 5-fold stratified cross-validation; stage 1, F1-macro=0.663 (SBERT+SMOTE+Random Forest); Stage 2, F1-Macro=0.966(TF-IDF+SMOTE+XGBoost); and stage 3, F1=0.9578 (TF-IDF+NumCat+RUS+Random Forest). In conclusión, this cascade design proved to be viable, modular, and interpretable for automating reviewer assignment using real-world data (213 articles, 117 evaluators). The sequential architecture allowed each stage to operate with the optimal representation for its data, contain errors, and validate the model with real-world data.

Keywords: machine learning (ML); peer review; artificial intelligence (AI); reviewer assignment; hybrid cascade models; natural language processing (NLP).

Resumen:

Las revistas científicas enfrentan demoras en las publicaciones de nuevos números, debido a la escasez de revisores expertos y el aumento de sometimiento de artículos. Este estudio propone una combinación de modelos con diseño en cascada que utilizó como técnicas de procesamiento del lenguaje natural (PLN) y aprendizaje automático (AA) para automatizar la asignación de revisores. Se adoptó un enfoque cuantitativo con diseño experimental sobre registros históricos extraídos de la plataforma OJS de una revista científica real del área de Ciencias Sociales centrado en las disciplinas Administración de Empresas, contabilidad, económicas, pedagogía aplicada al fomento empresarial, emprendimiento, desarrollo local, innovación, tecnología (2021-2025). El modelo se estructura en tres etapas secuenciales: clasificación de aprobación del resumen, clasificación del campo disciplinar y la clasificación de la rapidez del revisor. Los resultados fueron: se evaluaron 100 combinaciones de representaciones textuales, clasificadores y técnicas de balanceo, con validación cruzada estratificada de 5 folds; Etapa 1, F1-macro =0.663 (SBERT+SMOTE+Random Forest); Etapa 2, F1-macro =0.966(TF-IDF + SMOTE+XGBoost); y Etapa 3, F1= 0.9578 (TF-IDF + NumCat+ RUS+ Random Forest). En conclusión, este diseño en cascada demostró ser viable, modular e interpretable para automatizar la asignación de revisores con datos reales (213 artículos, 117 evaluadores). La arquitectura secuencial permitió que cada etapa opere con la representación óptima para sus datos, contener errores y validar el modelo con datos reales.

Palabras clave: aprendizaje automático (AA); revisión por pares; inteligencia artificial (IA); asignación de revisores; modelos híbridos en cascada; procesamiento del lenguaje natural (PLN).

INTRODUCTION

Scientific journals are responsible for publishing high-impact articles for the academic and scientific community. To meet this objective, editorial teams develop processes that involve not only the editorial committee but also require the participation of external reviewers considered experts in the thematic area of the manuscript, which indicates the importance of their role in the validation and quality assurance of published articles[1].

The traditional or semi-automated peer review system faces challenges at a global level; for example, the increase in article submissions creates an overload for editorial teams, making it difficult to identify suitable reviewers and compromising the quality of evaluations[2].

Moreover, these difficulties increase in Latin America, particularly in Colombia, where the scarcity of specialized reviewers, the lack of technological infrastructure and the low international visibility of journals worsen the problem [3], [4]; even factors such as language barriers, the concentration of local academic networks, the difficulty of identifying thematic affinities and delays in evaluation times have generated an increased risk of biases and conflicts of interest, reducing the transparency of editorial processes [5], [6], [7].

Fig. 1 shows the time taken by the peer review process.

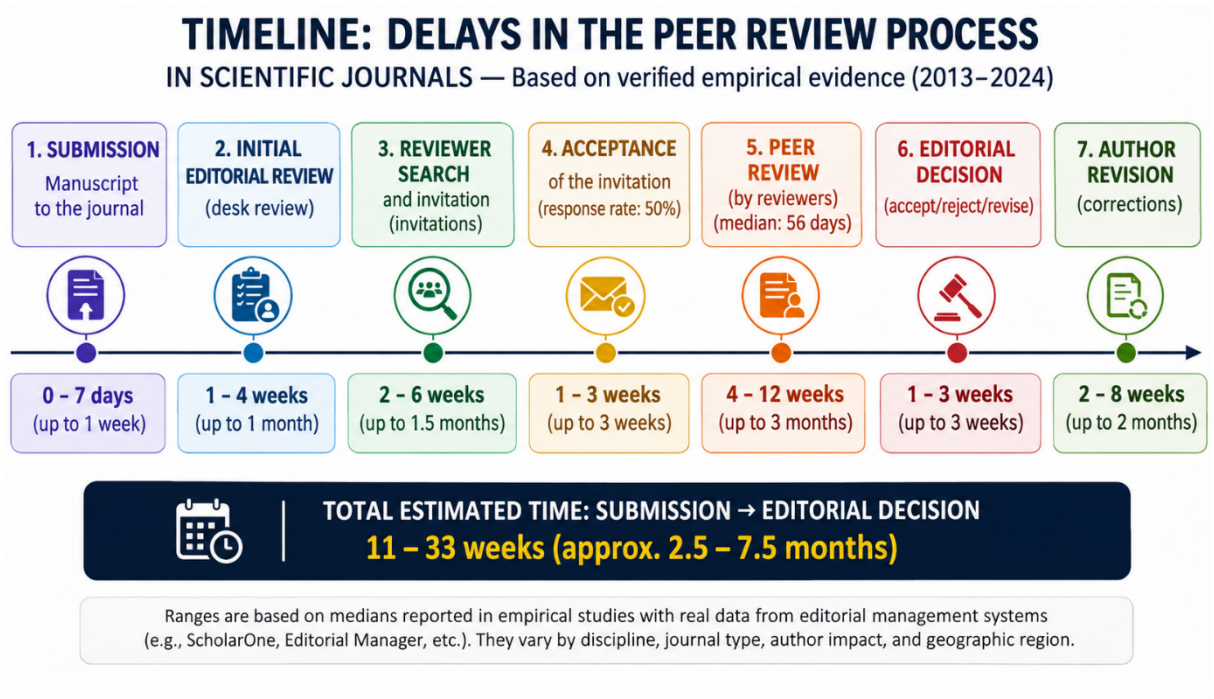

Fig. 1. *Timeline: peer review process.*
Source: prepared by the authors. Note: studies used to construct the timeline [8], [9], [10], [11], [12].

Regarding the automation of reviewer assignment, algorithms have been proposed that combine: references to aims and weight of publications to identify reviewers with thematic affinity[13]; approaches based on the depth and breadth of disciplinary knowledge, reducing barriers to locating experts[14]; systems such as PEERRec and PEERAssist that have explored the prediction of editorial decisions from interactions between manuscripts and reviewers[2],[15]; automated evaluations of review quality[16]; the analysis of linguistic and semantic features of articles[17]. All these solutions aim to generalize and strengthen the integrity of arbitration through the use of artificial intelligence (AI) and natural language processing (NLP) techniques, making it possible to transform articles into valid structures so that algorithms can support decision-making.

Although there are advances in this field, limitations persist in the integration of multiple sources of information, in architectures that integrate them sequentially, and in the precision of models, such as bibliometric metadata, academic profiles and textual content, which reduces their capacity to capture the multidimensional complexity of a reviewer's expertise [18], [19]. Second, the precision of current models is limited by the variability in similarity scores, as well as by the scarcity of quality labeled data that would allow an evaluation of the proposed algorithms [20], [21].

This research therefore proposes a hybrid cascade design that combines NLP and machine learning (ML) techniques for reviewer assignment in scientific journals. Its differential contribution relative to the studies analyzed here does not lie in the techniques employed (SBERT, TF-IDF, SMOTE, Random Forest, XGBoost or Logistic Regression), but in finding the

most optimal combination according to the nature of the data, since those studies were mainly oriented toward open-access databases derived from conferences such as ICLR and NeurIPS, and not toward scientific journals, which keep this information confidential. The design was implemented as a sequence of three stages and validated on real operational data within a Latin American editorial context with limited resources and scarce information, in order to improve the efficiency and quality of the peer review process.

THEORETICAL FRAMEWORK

Challenges in peer review and limitations in the manual process of editorial management

The peer review process is the main component for the validation of scientific knowledge[22]. However, it has faced challenges that compromise its efficiency, such as the increase in submitted manuscripts, the scarcity of reviewers and inconsistencies in evaluations[2],[23], generating overloads: approximately 20% of reviewers perform more than 70% of reviews globally[11]. In nursing journals, for example, between 25 and 30 invitations are required to achieve a single effective review[24]. The scarce recognition of reviewers, since it is a voluntary activity, the inappropriate use of AI tools and biases based on institutional affiliation, gender or nationality further deteriorate the system's trust[23],[25]. The study conducted by NeurIPS and cited in[2]documented that 57% of the articles accepted by one committee were rejected by others, and only 23% of reviewers state that they are completely confident in their recommendations[26].

These processes are carried out manually, worsening this situation[27], by limiting the capacity to process information efficiently and generating less accurate assignments[13],[28],[29]. Although editorial management systems exist, they do not capture the semantic complexity of scientific knowledge[30]. In Latin America and Colombia, where there is a lower density of researchers than in regions such as Europe or North America, this restricts the system's response capacity[3]. The concentration of scientific evaluation in certain geographic and linguistic contexts hinders access to international reviewers, increasing the burden on the few available bilingual specialists[6], who are called upon recurrently, with the consequent decrease in the quality of evaluations[4]. These practices tend to privilege reviewers from countries with greater scientific production, reducing the diversity of perspectives[31], and generating low international visibility of the journals, which restricts their positioning in global indexes[5],[32].

Artificial intelligence, semantic representation models and performance metrics in reviewer assignment systems

Artificial Intelligence (AI), supported by natural language processing (NLP) and machine learning (ML) techniques, emerges as an alternative to optimize multiple stages of the editorial process, allowing the extraction of semantic representations that facilitate comparison between manuscripts and reviewer profiles[22],[33],[34]. Based on the systematic literature review and application of prism methodology in Scopus, IEEE, and Springer databases with the keywords "Automated Peer Review Recommendation, Reviewer Assignment using NLP, AI for Academic Peer Review," 137 related articles were identified. After applying inclusion criteria (IA / ML / PLN applied, last 5 years, scarcity of reviewers, overload of evaluators) and exclusion (no empirical evidence, outside the central topic, no applicable proposal, debates without alternatives), 22 articles were analyzed, where 90% used NLP as the central component, highlighting pre-trained models such as BERT, SBERT, GPT, and BART, as well as semantic similarity techniques like TF-IDF, Word2Vec, and cosine similarity[2],[14],[23].

Vector representations of text and word embedding models such as Word2Vec have been widely used to measure thematic affinity[14],[35]. However, they have limitations when it comes to capturing the contextual complexity of texts. More recent models such as BERT and SBERT generate contextualized embeddings that substantially improve content interpretation[2],[36]. Likewise, the use of dynamic cosine similarity thresholds and the combination of quantitative metrics have been shown to improve accuracy in reviewer assignment[14],[35].

According to the literature, class balancing techniques should be in line with the characteristics of the data. These include random oversampling (ROS) and SMOTE, which seek to compensate for the underrepresentation of the minority class through duplication or synthetic generation of instances [37], [38]; random undersampling (RUS), which reduces the majority class [39]; and more recent generative approaches based on conditional adversarial networks (CTGAN), aimed at capturing complex distributions in mixed tabular data [40], [41]. The choice among these techniques usually depends on the sample size, the severity of the imbalance, and its interaction with the classifier used [42], [43].

Among the most widely used performance metrics are: F1 with 45%, in tasks with imbalanced classes such as distinguishing between high- and low-quality reviews, where overall accuracy would be misleading because it favors the majority class[16],[28],[44]; ROUGE with 27%, in text generation systems, allowing comparisons between different approaches[45],[46],[47]; precision and recall with 23%, in contexts where the interpretability of each component is a priority[13],[35]. Also included are the area under the curve (AUC-ROC), which validates the discriminative capacity of the model across different classification thresholds, and accuracy, which refers to overall performance on more balanced datasets, with 14%[23],[44],[48].

Design of hybrid cascade models for reviewer assignment and the state of knowledge compared with the study

Given the diversity of data types contained in the metadata captured by journals at the time of submitting an article, the performance of a single model integrating different tasks becomes complex and shows low performance[47]. For this reason, hybrid cascade models stand out as a strategy for dividing the complexity of a model into several models that operate sequentially. These models can address issues related to reviewer assignment, structured in several stages: (i) preprocessing and semantic representation of the manuscript, using techniques such as SBERT or BERT; (ii) extraction of relevant features, including topics, keywords or specific content aspects; (iii) calculation of similarity or affinity between the article and reviewer profiles, and (iv) classification or ranking of reviewers[49]. In this architecture, the output of each stage constitutes the input of the next, which distinguishes it from end-to-end architectures, where all stages are optimized jointly[21],[50].

The studies reviewed have used the ICLR or NeurIPS conferences as their database, which consist of open peer review and abundant information[2],[23], which limits their applicability in small or medium-sized journals with restricted resources. Therefore, this work proposes an automatic reviewer assignment system through a model with a hybrid cascade design that incorporates NLP and ML techniques, adapted to a real editorial process.

METHODOLOGY

Research approach and design

The research adopted a quantitative approach with an experimental design aimed at building and validating a hybrid cascade model, based on NLP and ML for automating reviewer assignment. This approach is consistent with proposed research on automating the peer review process, which is based on predictive models and quantitative evaluations[2],[13],[23].

In this sense, the model developed was a hybrid cascade model, following principles of modular design and architectural patterns for machine learning systems that favor reproducibility, interpretability and scalability[51]. The system's performance was evaluated using metrics that ensure the validity and reliability of the research results.

Data collection and preprocessing

The data were extracted from the Open Journal Systems (OJS) platform, version 3.3.0.6, corresponding to the period 2021 to 2025, obtaining 221 initial records. The variables collected were grouped into two main datasets: one focused on the characteristics of the articles and the other on those of the reviewers, in order to facilitate the differentiated analysis of each dimension of the tasks.

Given the scope of the journal and the characteristics of the articles submitted for evaluation and possible publication, the areas of knowledge were reclassified according to the OECD Frascati Manual and the Publindex 2026 classification and recognition model for national scientific journals, consolidating into 6 categories: Social Sciences, Education, Economic, Administrative and Accounting Sciences, Agricultural and Veterinary Sciences, and Exact Sciences. This reclassification was carried out because the areas of interest originally registered in the editorial system did not follow a standardized taxonomy, which hindered their direct use as a classification variable.

However, the use of this taxonomy could introduce a possible granularity bias with respect to areas of knowledge with diffuse disciplinary boundaries: the journal analyzed is oriented toward the Social Sciences, with a marked concentration in the area of Economic and Accounting Sciences (56%) and a considerably smaller representation of the other areas (44%), which may have artificially favored the lexical discriminability observed in stage 2 by benefiting the separation of the majority class from the minority class. The target variable of stage 1 presented an imbalance: 90% approved articles and 10% not approved. This behavior has also been reported in similar studies on editorial prediction in academic contexts[2],[44].

Preprocessing included: (1) cleaning and removal of incomplete and duplicate records and of recommendations canceled by the reviewers; (2) normalization of categorical variables through string standardization; (3) transformation of temporal variables for the calculation of response time and compliance; and (4) processing of textual data (abstracts, areas of interest and recommendations) through NLP techniques, such as text cleaning, removal of special characters, tokenization, lemmatization and vectorization with multiple representation techniques, with the purpose of converting unstructured information into numerical representations suitable for statistical analysis and model training. After this process, 213 records were obtained, divided into two datasets: (1) articles(213 records, 6 areas of interest according to the Publindex 2026 Classification and Recognition Model for National Scientific Journals)and (2) reviewers (117 unique records), whose distribution is presented in Table 1.

To ensure the model can be reproduced, Table 2 summarizes each stage, the input variables used and the corresponding target variable. All variables were taken directly from the historical records of the OJS system and were normalized for stage 3 using StandardScaler (numerical variables) and OneHotEncoder (categorical variables), fitted on the training subset of each experimental partition.

Table 1: Distribution of reviewers by MinCiencias Publindex 2026 areas of interest.

Areas of interest (Minciencias Publindex 2026)	No. of Reviewers	% of total
Economic, Administrative and Accounting Sciences	72	61.5%
Education.	28	23.9%
Agricultural and Veterinary Sciences	3	2.6%
Social Sciences	12	10.3%
Exact Sciences	2	1.7%
Total	117	100%

Sources: project authors

Table 2: input and target variables for each stage.

Stage	Input variable	Target variable
E1-abstract acceptance classification	Textual abstract of the article, vectorized using the optimal textual representation	Recommendation: Accepted (1) / Not accepted (0)
E2- disciplinary field classification	Textual abstract of the article, vectorized using the optimal textual representation	Disciplinary field: Economics, Administrative and Accounting Sciences (1) / Other areas (0)
E3 – reviewer classification	(i) Disciplinary area of the reviewer vectorized using textual representation; (ii) 8 numerical variables scaled with StandardScaler, Google Scholar h-index, average days of assignment, No. of historical reviews, % of reviews submitted before the deadline and 4 temporal dispersion statistics (maximum, minimum, standard deviation and median days relative to the deadline); (iii) 2 nominal categorical variables (education level, nationalities encoded with OneHotEncoder)	Reviewer speed: Fast (1) / Not fast (0) constructed from the 33rd percentile of the average days relative to the submission deadline.

Handling of class imbalance, data splitting and validation strategy

The datasets used in this study showed a severe imbalance at stage 1 (approval: ratio 9.1:1) and a mild one at stage 2 (disciplinary field: 1.3:1); for reviewers the ratio was approximately 2:1. According to[54], a dataset is imbalanced when one or several classes are represented less frequently than others, which generates biased predictions; techniques at the data, algorithm and evaluation levels exist to mitigate this.

Four balancing techniques were implemented: ROS (Random Oversampling), which duplicates existing examples of the minority class without generating new information[37]; SMOTE, which generates synthetic examples by interpolating nearby samples of the minority class[38], with k_neighbors=2 for article classification and k_neighbors=3 for reviewers; RUS (Random Undersampling), which reduces the majority class by randomly removing samples[39]; and CTGAN (Conditional Tabular Generative Adversarial Network), a conditional generative adversarial network that produces synthetic tabular data by capturing complex distributions in numerical and categorical variables[40],[41]. The most suitable technique was selected empirically by evaluating each technique in combination with the different textual representations and classifiers of the model[42],[43], using evaluation metrics such as accuracy, precision, recall and F1-macro[44],[48].

In the article classification stage, stratified 5-fold cross-validation was used on the 213 records, without a prior fixed train/test split, so all the metrics reported correspond to the mean and standard deviation of the five folds, ensuring that the class proportion remained representative in each iteration [55]. The TF-IDF and Word2Vec representations were fitted only on the training data of each fold; the pre-trained embeddings (BERT, SBERT, Gemini) were generated with caching before the partition, without introducing leakage [56]. The metric used was F1-macro, which penalized collapse onto the minority class compared with F1-binary [57]. For the reviewer classification model, a stratified 80/20 partition (train/test) was assumed, with 5-fold cross-validation on the training subset. Hyperparameter tuning was applied, and the search was performed using gridSearchCV with an internal 3-fold StratifiedKfold as the criterion. The metrics used were: F1-weighted on the external test, precision, recall and AUC-ROC. The target variable (Fast=1 / Not fast = 0) was constructed from the 33rd percentile of the average days relative to the submission deadline.

To ensure the reproducibility of the experiments, a common random seed (random_state = 42) was set in all splits of the dataset, as well as in the classification algorithms used. Class balancing was applied only to the training subset of each fold, keeping the test subset free of contamination by synthetic data[58].

Textual representation techniques and classifiers

The texts were converted into numerical variables using the vectorization techniques described in Table 3.

Table 3: Textual representation techniques used.

Technique	Description	Application
TF-IDF	Term Frequency-Inverse Document Frequency; vectorization (1-2 n-grams, 5000 features)	Article and reviewer classification
Word2Vec	Distributed embedding trained on the training corpus of each fold (size = 300, window = 5, epochs = 30)	
BERT	Bert-base-multilingual-cased; 768-dimensional contextual [CLS] representation, with disk cache	
SBERT	Paraphrase-multilingual-MiniLM-Ll2-V2; sentence embeddings for semantic similarity.	
Gemini Emb. 2	Models/Gemini-embedding-2-preview; 768-dimensional output, task type CLASSIFICATION	

For the reviewer classification model, eight numerical variables were integrated (Google h-index, assignment days, historical evaluations and days relative to the deadline), standardized with StandardScaler, and two nominal categorical variables

(Education Level and Nationality) encoded with OneHotEncoder to avoid artificial numerical relationships[59]. The text, numerical and categorical features were concatenated horizontally before training. Four classifiers were evaluated with hyperparameter tuning via GridSearchCV, summarized in Table 4.

Table 4: Classifiers and hyperparameter configuration.

Classifier	Main configuration	Fitting
Logistic Regression	solver= 'saga', class_weight= 'balanced', max_iter=2000	GridSearchCV(C∈{0.1,1,5,10})
SVM	Kernel= ' rbf', probability =true, class_weight= 'balanced'	GridSearchCV (C, gamma)
Random Forest	N_estimators = 200-300, class_weight= 'balanced'	GridSearchCV(n_estimator, max_depth, learning_rate)
XGBoost	Scale_pos_weight; computed on the original y_train of the fold; eval_metric='logloss '	GridSearchCV(n_estimators, max_depth, learning_rate)

Architecture of the hybrid cascade design

The model architecture is framed by three sequentially chained binary classification stages, where the output of each stage conditions the activation of the next (Fig. 2).

In stage 1 (abstract acceptance classification), the article's textual abstract was vectorized and classified as Accepted (1) and Not accepted (0), determining whether the article continues in the system or receives an immediate response (see Table 2).

In stage 2 (disciplinary field classification), the abstract of the article accepted in the previous stage was vectorized and classified into the disciplinary field Economics, Administrative and Accounting Sciences (1) and other areas (0), acting as a thematic filter that delimits the subset of candidate reviewers in the following stage (see Table 2).

In stage 3 (reviewer classification), a feature vector was built for each reviewer from three sources of information: textual representations of thematic keywords and nominal categorical numeric variables (see Table 2). With these, response speed was classified (Fast (1) / Not fast (0)), constructed from the 33rd percentile of the average number of days relative to the submission deadline. The model's posterior probability p(Fast) generated a ranking of the five reviewers with the highest probability of timely response. For each of the three stages, the set of textual representations, balancing techniques and classifiers described in section 3.4 was evaluated independently, selecting in each case the combination with the best performance according to the nature of the data in each stage; the results of this comparison and the final configuration selected are presented in section 4.4.

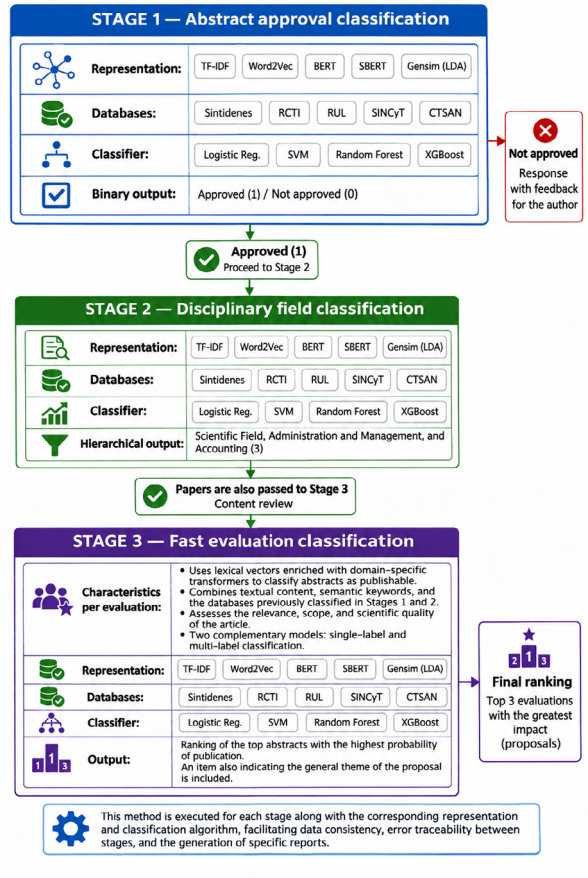

Fig. 2. *Hybrid cascade architecture.*
Source: project authors.

The predictor variables selected for Stage 3 are imperfect proxies of the reviewer's actual expertise and availability, and should be interpreted with caution. The Google Scholar h-index, although widely available, does not directly reflect the reviewer's thematic specialty: a low h-index does not necessarily imply lower expertise, particularly in highly specialized areas of knowledge or where scientific output is lower but thematic relevance may be high. Similarly, the historical response times used as a predictor variable show considerable variability and are subject to external factors not controlled by the model, such as the reviewer's workload, their time availability, institutional commitments and academic seasonality (for example, vacation periods or end of semester). These limitations do not invalidate the use of such variables, but they condition the interpretive scope of the predictions in Stage 3 and must be taken into account when deploying the model in a production environment.

Fig. 2 summarizes the different combinations made in each of the stages described.

RESULTS

Characteristics of the datasets

Dataset 1 of articles is intended for stages 1 (E1) and 2 (E2), with 213 unique records and four variables: identifier, abstract, area of interest and editorial recommendation. Dataset 2 of reviewers, for stage 3 (E3), with 213 reviewer records consolidated into 117 unique reviewers with 11 original variables expanded to 16 after preprocessing. The binary target variable of E3 was built with the 33rd percentile (threshold = -1.0 days): a reviewer is classified as Fast (1) if their average days relative to the limit is ≤ -1.0, indicating early submission.

Experimentation in Stages 1 and 2

Building the model involved exploring a search space of 100 unique configurations per stage (5 representations x 4 classifiers x 5 balancing strategies) on the articles dataset, using F1-macro as the main metric with an automatic collapse detector: F1-min < 0.10 in the minority class in all folds, the combination was excluded from the ranking.

In Stage 1 (Abstract acceptance classification) (with severe imbalance 9.1:1) (E1): A total of 5 combinations with total collapse in all 5 folds were generated, all involving TGAN + Logistic Regression or SVM. The most optimal combination was SBERT+SMOTE+Random Forest with F1-macro = 0.663±0.122, followed by TF-IDF+SMOTE+Random Forest (0.659) and TF-IDF+SMOTE+Logistic Regression (0.656). The standard deviation value demonstrated the difficulty of the problem: 21 negative instances, presenting between 3 and 4 examples of the minority class in test. The performance of the accumulated global classifier was macro precision = 0.63, macro recall = 0.63, accuracy = 87%. This difference between the high accuracy and the considerably lower precision and recall values demonstrates the effect of class imbalance (90% accepted versus 10% not accepted): accuracy is favored by the majority class and is uninformative about the model's real capacity to identify non-accepted articles, whereas the macro metrics, by weighting both classes equally, reflect the difficulty of the problem more faithfully. For this reason F1-macro was chosen as the main metric for model selection at this stage. A similar, though less pronounced, pattern is observed in the distribution of areas of interest (56% Economics, Administrative and Accounting Sciences, 44% other areas), which may have slightly favored the performance of the combinations evaluated in stage 2. Table 5 shows the ranking of the best combinations.

Table 5: Top 6 combinations Stage 1 (F1-macro, without total collapse)

Embedding	Balancing	Classifier	F1-macro	F1 macro std
SBERT	SMOTE	Random Forest	0.6633	0.1220
TF-IDF	SMOTE	Random Forest	0.6586	0.1248
TF-IDF	SMOTE	Logistic regression	0.6560	0.1295
BERT	No balancing	Random Forest	0.6452	0.1170
Gemini Emb. 2	No balancing	Random Forest	0.6452	0.1170
Word2Vec	No balancing	Random Forest	0.6452	0.1170

Note. The row in green: combination selected for the cascade model. Source: Own elaboration.

Table 6 compares F1 by balancing technique in both stages. ROS obtained the highest F1 in E1 (0.6011), followed by SMOTE (0.5948), both higher than without balancing (0.5848). TGAN showed a higher frequency of collapse (2.65 folds on average), and RUS reduced the recall of the majority class. The difference between ROS (0.6011) and SMOTE (0.5948) amounts to 0.0063 in F1-macro, smaller than the standard deviation between folds observed for both techniques, so on its own it is not statistically conclusive. It should also be noted that, in the collapsed folds column of Table 5, ROS showed a slightly lower average than SMOTE in Stage 1; in this sense, the choice of SMOTE rests not on empirical superiority over ROS in these two specific indicators, but on a theoretical criterion of generalization: whereas ROS duplicates existing examples of the minority class without adding variability, which increases the risk of overfitting to repeated instances, SMOTE generates synthetic examples by interpolating between nearby neighbors, favoring a smoother decision boundary that is potentially more generalizable outside the cross-validation sample used. Since the numerical difference between the two techniques is marginal, it is acknowledged that ROS could be an equally valid alternative for validations with larger datasets.

Table 6: Average F1-macro by balancing techniques (E1 and E2, 100 combinations).

Balancing	F1-macro E1	F1-macro E2	Collapsed folds
ROS	0.6011	0.8718	0.85
SMOTE	0.5948	0.8717	0.90
No balancing	0.5848	0.8752	1.00
TGAN	0.5632	0.8748	2.65
RUS	0.3156	0.8528	0.55

Source. Own elaboration

Stage 2 (Classification disciplinary field) (with slight imbalance 1.3:1) (E2): It obtained the most balanced set (120 vs. 93 records), allowing all F1-macro combinations to be >0.90 without collapse. The combination TF-IDF + SMOTE+XGBoost reached F1-macro = 0.9662±0.0363, with an AUC-ROC =0.9915 and F1-min_class =0.9615. Folds 2 and 3 achieved F1=1.0000, indicating the high lexical discriminability of the economic-administrative vocabulary in the spaCy lemmatization. Table 7 shows the best combinations in this stage.

Table 7: Top5 combinations Stage 2 (F1-macro, 5 folds CV)

Emb	Bal	Classif	F1-macro	F1 std
TF-IDF	SMOTE	XGBoost	0.9662	0.0363
Gemini Emb2	No balancing	XGBoost	0.9621	0.0322
SBERT	No balancing	Random Forest	0.9613	0.0247
Word2Vec	ROS	XGBoost	0.9516	0.0264
BERT	SMOTE	XGBoost	0.9466	0.0446

Source. Own elaboration. Note. Emb=Embeddings, Bal=balancing, Clasif=classifier; the green shading is the combination selected for the cascade design.

Experimentation in Stage 3: Reviewer classification (E3)

Five representation techniques combined with 4 classifiers were evaluated on 117 unique reviewers, with an 80/20 stratified split (train =93, test=24). The small size of the set reflects the overall limitation of the dataset (213 records in Stages 1 and 2, 117 reviewers and an external test set of only n=24 in Stage 3), which conditions the statistical robustness and the generalization of the metrics reported in the stage, whose reliability is sensitive to the size of the test set; the results obtained in this stage must be interpreted as samples of a proof of concept and not as an estimate of stable performance in production. Likewise, 4 resampling balancing techniques (ROS; SMOTE, RUS and CTGAN) were evaluated for each of the textual representations, instead of applying a single fixed strategy a priori. This comparison showed that the optimal technique varies according to the representation: RUS was preferable for TF-IDF+ NumCat, SBERT+ NumCat, BERT+NumCat and Gemini+NumCat, whereas CTGAN was more suitable for Word2Vec+NumCat. Undersampling resampling (RUS) therefore tends to favor most of the representations evaluated in this dataset, although this relationship should be confirmed in future studies with a larger dataset[63].

Two points are important to note. First, the 33rd percentile threshold (p33) used to create the target variable (Fast/not fast) was calculated on the complete dataset, when it should have been limited to the training subset in each experimental partition; this decision introduces a partial information leakage in the definition of the target variable, since the threshold incorporates information from the test set, which may partly compromise the validity of the reported evaluation. The complete correction of this procedure would involve repeating the experimentation of Stage 3 with the threshold recalculated per partition; the conclusions derived from this stage must be interpreted with a limited inferential scope, as a preliminary result of technical feasibility and not as an unbiased estimate of the model's performance in production. Second, for Gemini Embedding 2, 256-dimensional representations were used instead of the 768 that are usual for the model, due to restrictions of the API used; this dimensional reduction established an experimental limitation that affected the comparability of Gemini against the other representations evaluated and may have biased the results, disadvantaging its relative position within the performance ranking of Table 8.

Table 8: Comparison of the techniques at stage 3 (external test set, n = 24)

Technique	Bal.	Clasif	F1-w	AUC	CV F1
TF-IDF + NumCat	RUS	RF	0.9578	0.9926	0.9475±0.0471
BERT + NumCat	CTGAN	RL	0.9179	0.8741	0.8507±0.068
Word2Vec+NumCat	RUS	RF	0.9141	0.9704	0.9570±0.039
SBERT+NumCat	RUS	LR	0.8761	0.9185	0.8301±0.076
Gemini+NumCat	RUS	LR	0.8761	0.8963	0.8518±0.084
Ensemble(soft)	SMOTE	RF+XGB+SVM+LR	0.8733	0.9556	0.9047±0.0513

Note. LR=Logistic Regression, RF=Random Forest, Bal. = balancing, CV F1=cross-validation on the training subset (5 folds). The light blue row is the selected combination of the most optimal model for that stage (TF-IDF+NumCat with RUS and Random Forest, F1=0.9578). The combinations SBERT+NumCat and Gemini+NumCat, both with RUS and Logistic Regression, achieved exactly the same performance on the test set; given the small size of the test set (n=24), this coincidence does not correspond to a transcription error, but rather to the fact that both models correctly classified the same subset of reviewers.

The combination TF-IDF + NumCat + RUS + Random Forest was the most optimal, with F1 = 0.957 (precision = 0.9609, recall = 0.9583), achieving good stability in cross-validation (CV F1 = 0.9475 ± 0.0471) relative to the rest of the combinations

evaluated. It was followed by BERT + NumCat CTGAN with Logistic Regression (F1 = 0.9179) and Word2Vec + NumCat with ROS and Random Forest (F1 = 0.9141). SBERT + NumCat and Gemini + NumCat, both with RUS and Logistic Regression, reached intermediate performance and were exactly identical to each other (F1 = 0.8761), which is consistent with the limitation already noted regarding the small size of the test set (n = 24). The Soft Voting Ensemble (RF + XGB + SVM + LR) obtained F1 = 0.873, without providing any improvement over the selected model. Table 8 compares the main combinations evaluated. The combination TF-IDF + NumCat + RUS + Random Forest was the most optimal, with F1 = 0.957 (precision = 0.9609, recall = 0.9583), achieving good stability in cross-validation (CV F1 = 0.9475 ± 0.0471) relative to the rest of the combinations evaluated. It was followed by BERT + NumCat CTGAN with Logistic Regression (F1 = 0.9179) and Word2Vec + NumCat with ROS and Random Forest (F1 = 0.9141). SBERT + NumCat and Gemini + NumCat, both with RUS and Logistic Regression, reached intermediate performance and were exactly identical to each other (F1 = 0.8761), which is consistent with the limitation already noted regarding the small size of the test set (n = 24). The Soft Voting Ensemble (RF + XGB + SVM + LR) obtained F1 = 0.873, without providing any improvement over the selected model. Table 8 compares the main combinations evaluated.

Performance of the cascade system

Table 9 summarizes the combinations chosen for each stage, with the highest F1-macro (E1 and E2) and F1-weighted (E3) predominating and no collapse onto the minority class, which ensures that the system does not drift toward a single outcome for rejected articles or slow reviewers.

Table 9: Combinations selected for the hybrid cascade system.

Stage	Optimal combination	Metric	Validation
1	SBERT+SMOTE+Random Forest	F1-macro: 0.663±0.122	stratified CV-5
2	TF-IDF+SMOTE+XGBoost	F1-macro: 0.966±0.036	Stratified CV-5
3	TF-IDF+NumCat+RUS+Random Forest	F1: 0.9578	80/20 external

Source. Own elaboration. Note. CV-5= 5-fold stratified cross-validation, 80/20= external partition.

This cascade design allowed each stage to operate with the representation and algorithm best suited to the nature of the data, avoiding the transfer of errors between stages through explicit binary decision thresholds[62]. Fig. 3 summarizes the data set processing for each model generated at each stage.

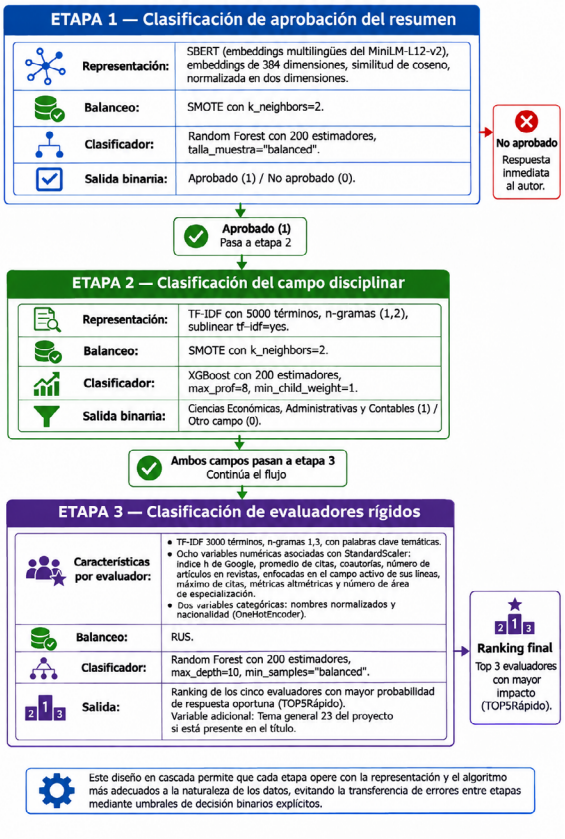

Fig 3. Architecture of the hybrid cascade design with the results at each of the stages.
Source. project authors.

The test was carried out with real articles submitted to the target scientific journal, selecting those that contained contradictory reviewer decisions. In this test, the posterior probability estimated by the E1 classifier for the "Approved" class was 75.8%; this figure corresponds to the probabilistic output of the model (not to the accuracy measure nor to a statistical confidence interval) and should be interpreted as the classifier's degree of certainty about that specific case, within the framework of the overall performance reported for Stage 1 (F1_macro=0.663). Similarly, E2 assigned the disciplinary field a posterior probability of 54.8% toward the Engineering field, and E3 generated a ranking of 5 possible reviewers according to their estimated probability of timely response, as shown in Fig. 4. Given the operational applicability of the system, its functional conditions must be specified. The model depends on historical variables (number of previous reviews, past response time) that are not available for new reviewers with no history in the system; in these cases, the Stage 3 model will not be able to generate a reliable prediction and will need a fallback rule (for example, random assignment weighted by thematic affinity) until sufficient history is accumulated. Likewise, in scenarios of editorial overload, peaks in article submissions and reviewer ranking produced by the model will not dynamically incorporate the reviewer's current availability, so its use must be complemented with a manual availability check before the formal invitation. Therefore, the system can be considered functional as a decision-support tool for editorial decisions in the experimental context evaluated (213 articles, 117 reviewers from a real journal), but its deployment in production at a larger scale will require additional validation with new reviewers and under conditions of variable load.



Fig 4. Graphical interface of the hybrid cascade system.
Source. Own elaboration.

CONCLUSIONS

The hybrid cascade design built in this study proved to have functional performance for automatic reviewer assignment under the conditions evaluated in a small dataset (213 articles, 117 reviewers), with stratified cross-validation of 5 folds in stages 1 and 2, and the external partition of only 24 cases in Stage 3. In the limitations and justifications noted regarding the calculation of the p 33 threshold, the sample size and the proxy nature of variables such as the h-index, where the architecture allowed assigning optimal representations per subproblem, containing errors between stages and validating the model with real operational data, without implying a guarantee of equivalent performance in larger-scale production scenarios or with new reviewers. The selection of textual representations, balancing techniques and classifiers per stage was decisive: SBERT proved optimal in E1 due to its ability to capture semantic similarity of sentences; TF-IDF with lemmatization dominated in E2 due to the lexical discriminability of the vocabulary; and Random Forest in E3 in testing, demonstrating that there is no universally superior representation[64], but rather that everything depends on the nature of the data of each task. Additionally, with the severe imbalance in the E1 articles dataset (9.1:1), the SMOTE(K=2) class balancing was more stable than CTGAN (2.65 collapsed folds) and RUS, while for high-dimensional embeddings it proved more viable in the hyperparameter search. Lemmatization with spaCy and the adjustment of the vectorizer on the training data were necessary conditions for reliable generalization estimates. The thematically enriched textual profile of E3 showed that the semantic design of the dataset can compensate for sample size limitations in sparse representations[58].

Limitations

The dataset used in this study is small, which produced high inter-fold variance in E1 (21 negative instances); the p33 threshold of E3 was calculated on the complete dataset and in production must be restricted to the training subset; and Gemini Embedding 2 was generated with 256-d (vs. the usual 768-d) due to API restrictions, potentially affecting its position in the ranking.

Future work

Implement fine-tuning of multilingual BERT on the corpus of Spanish abstracts to improve the F1-macro of E1; expand the dataset with other journals to extend disciplinary coverage; implement incremental updating of the E3 model to mitigate concept drift in the reviewer history; and apply LLM evaluation as zero-shot classifiers in E1 for contexts with very scarce

data. On the other hand, actions such as the following should be implemented: recalculate the p 33 threshold only on the training subset in each partition to eliminate the bias currently acknowledged in Stage 3; deepen the selection and validation of alternative predictive variables to the proxies used (h-index, historical response times); evaluate the temporal generalization of the model through validation with data from periods subsequent to the training period (2021-2025); and explore mechanisms that allow assigning new reviewers without prior history to adapt the ranking to real availability in editorial overload scenarios.

REFERENCES

[1] J. P. Gisbert y M. Chaparro, «Reglas y consejos para ser un buen revisor por pares de manuscritos científicos», Gastroenterol. Hepatol., vol. 46, n.o 3, pp. 215-235, mar. 2023, doi: 10.1016/j.gastrohep.2022.03.005.

[2] P. K. Bharti, T. Ghosal, M. Agarwal, y A. Ekbal, «PEERRec: An AI-based approach to automatically generate recommendations and predict decisions in peer review», Int. J. Digit. Libr., vol. 25, n.o 1, pp. 55-72, mar. 2024, doi: 10.1007/s00799-023-00375-0.

[3] J. A. Huete-Perez y N. Salvatierra, «Assessing biomedical research capacities in selected countries of Latin America: challenges, opportunities, and recommendations», Front. Res. Metr. Anal., vol. 10, jul. 2025, doi: 10.3389/frma.2025.1594303.

[4] A. Severin y J. Chataway, «Overburdening of peer reviewers: A multi-stakeholder perspective on causes and effects», Learn. Publ., vol. 34, n.o 4, pp. 537-546, 2021, doi: 10.1002/leap.1392.

[5] A. Becerril García y S. Córdoba González, Eds., Conocimiento abierto en América Latina: trayectoria y desafíos. en Colección Grupos de trabajo. Erscheinungsort nicht ermittelbar: CLACSO, 2021.

[6] J. A. C. Osorio, «Problemáticas de las revistas científicas colombianas»:, Sci. Tech., vol. 29, n.o 4, pp. 145-147, dic. 2024, doi: 10.22517/23447214.25746.

[7] J. C. Calderón, «Dear Editor», Colomb. Medica, vol. 42, n.o 3, pp. 406-407, 2011, doi: 10.25100/cm.v42i3.889.

[8] N. Felisi, R. Vecchio, y A. Odone, «When peer review drags on: the harm to early career researchers», Front. Res. Metr. Anal., vol. 11, feb. 2026, doi: 10.3389/frma.2026.1740381.

[9] B.-C. Björk y D. Solomon, «The publishing delay in scholarly peer-reviewed journals», J. Informetr., vol. 7, n.o 4, pp. 914-923, oct. 2013, doi: 10.1016/j.joi.2013.09.001.

[10] J. Huisman y J. Smits, «Duration and quality of the peer review process: the author's perspective», Scientometrics, vol. 113, n.o 1, pp. 633-650, oct. 2017, doi: 10.1007/s11192-017-2310-5.

[11] Publons, «Publons' Global State Of Peer Review 2018», Publons, London, UK, sep. 2018. doi: 10.14322/publons.GSPR2018.

[12] B. Aczel, B. Szaszi, y A. O. Holcombe, «A billion-dollar donation: estimating the cost of researchers' time spent on peer review», Res. Integr. Peer Rev., vol. 6, n.o 1, p. 14, nov. 2021, doi: 10.1186/s41073-021-00118-2.

[13] T. A. Mahmud, B. M. M. Hossain, y D. Ara, «Automatic Reviewers Assignment to a Research Paper Based on Allied References and Publications Weight», en 2018 4th International Conference on Computing Communication and Automation (ICCCA), Greater Noida, India: IEEE, dic. 2018, pp. 1-5. doi: 10.1109/CCAA.2018.8777730.

[14] X. Zheng, P. Li, y X. Wu, «An Automatic Paper-Reviewer Recommendation Algorithm Based on Depth and Breadth», IEEE Trans. Emerg. Top. Comput. Intell., vol. 9, n.o 1, pp. 607-616, feb. 2025, doi: 10.1109/TETCI.2024.3402694.

[15] P. K. Bharti, S. Ranjan, T. Ghosal, M. Agrawal, y A. Ekbal, «PEERAssist: Leveraging on Paper-Review Interactions to Predict Peer Review Decisions», en Towards Open and Trustworthy Digital Societies, Springer, Cham, 2021, pp. 421-435. doi: 10.1007/978-3-030-91669-5_33.

[16] L. Ramachandran, E. F. Gehringer, y R. K. Yadav, «Automated Assessment of the Quality of Peer Reviews using Natural Language Processing Techniques», Int. J. Artif. Intell. Educ., vol. 27, n.o 3, pp. 534-581, sep. 2017, doi: 10.1007/s40593-016-0132-x.

[17] A. Perković Paloš et al., «Linguistic and semantic characteristics of articles and peer review reports in Social Sciences and Medical and Health Sciences: analysis of articles published in Open Research Central», Scientometrics, vol. 128, n.o 8, pp. 4707-4729, ago. 2023, doi: 10.1007/s11192-023-04771-w.

[18] J. Jovanovic y E. Bagheri, «Reviewer assignment problem: A scoping review», 13 de mayo de 2023, arXiv: arXiv:2305.07887. doi: 10.48550/arXiv.2305.07887.

[19] X. Zhao y Y. Zhang, «Reviewer assignment algorithms for peer review automation: A survey», Inf. Process. Manag., vol. 59, n.o 5, p. 103028, sep. 2022, doi: 10.1016/j.ipm.2022.103028.

[20] C. Cousins, J. Payan, y Y. Zick, «Into the Unknown: Assigning Reviewers to Papers with Uncertain Affinities», arXiv.org. Accedido: 4 de mayo de 2026. [En línea]. Disponible en: https://arxiv.org/abs/2301.10816v2

[21] D. Mimno y A. McCallum, «Expertise modeling for matching papers with reviewers», en Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, en KDD '07. New York, NY, USA: Association for Computing Machinery, ago. 2007, pp. 500-509. doi: 10.1145/1281192.1281247.

[22] H. Bauchner y F. P. Rivara, «Use of artificial intelligence and the future of peer review», Health Aff. Sch., vol. 2, n.o 5, p. qxae058, may 2024, doi: 10.1093/haschl/qxae058.

[23] I. Boukhris y C. Zaâbi, «A GAN-BERT based decision making approach in peer review», Soc. Netw. Anal. Min., vol. 14, n.o 1, p. 107, may 2024, doi: 10.1007/s13278-024-01269-y.

[24] S. M. Kadri, N. Dorri, M. Osaiweran, P. Garyali, y M. Petkovic, «Scientific Peer Review in an Era of Artificial Intelligence», en Scientific Publishing Ecosystem, Springer, Singapore, 2024, pp. 397-413. doi: 10.1007/978-981-97-4060-4_23.

[25] T. Ghosal, S. Kumar, P. K. Bharti, y A. Ekbal, «Peer review analyze: A novel benchmark resource for computational analysis of peer reviews», PLOS ONE, vol. 17, n.o 1, p. e0259238, ene. 2022, doi: 10.1371/journal.pone.0259238.

[26] A. Barnett, L. Allen, A. Aldcroft, T. L. Lash, y V. McCreanor, «Examining uncertainty in journal peer reviewers' recommendations: a cross-sectional study», R. Soc. Open Sci., vol. 11, n.o 9, p. 240612, doi: 10.1098/rsos.240612.

[27] W. Yuan, P. Liu, y G. Neubig, «Can We Automate Scientific Reviewing?», 30 de enero de 2021, arXiv: arXiv:2102.00176. doi: 10.48550/arXiv.2102.00176.

[28] R. Verma, K. Shinde, H. Arora, y T. Ghosal, «Attend to Your Review: A Deep Neural Network to Extract Aspects from Peer Reviews», en Neural Information Processing, Springer, Cham, 2021, pp. 761-768. doi: 10.1007/978-3-030-92310-5_88.

[29] C. Candal-Pedreira, J. Rey-Brandariz, L. Varela-Lema, M. Pérez-Ríos, y A. Ruano-Ravina, «Challenges in peer review: how to guarantee the quality and transparency of the editorial process in scientific journals», An. Pediatría Engl. Ed., vol. 99, n.o 1, pp. 54-59, jul. 2023, doi: 10.1016/j.anpede.2023.05.006.

[30] S. P. J. M. (. S. Horbach y W. (. W. Halffman, «The changing forms and expectations of peer review», Res. Integr. Peer Rev., vol. 3, n.o 1, p. 8, sep. 2018, doi: 10.1186/s41073-018-0051-5.

[31] J. P. Tennant et al., «A multi-disciplinary perspective on emergent and future innovations in peer review», 29 de noviembre de 2017, F1000Research. doi: 10.12688/f1000research.12037.3.

[32] M. Salatino y G. Banzato, «Confines históricos del Acceso Abierto latinoamericano», Conoc. Abierto En América Lat. Trayectoria Desafíos Ciudad Autónoma B. Aires CLACSO Univ. Autónoma Estado México 2021 P 79-115, 2021, Accedido: 19 de septiembre de 2025. [En línea]. Disponible en: https://www.memoria.fahce.unlp.edu.ar/library?a=d&c=libros&d=Jpm4982

[33] M. Majadly y M. Last, «Leveraging peer-review aspects for extractive and abstractive summarization of scientific articles», Int. J. Data Sci. Anal., vol. 19, n.o 3, pp. 537-550, abr. 2025, doi: 10.1007/s41060-024-00665-z.

[34] A. Checco, L. Bracciale, P. Loreti, S. Pinfield, y G. Bianchi, «AI-assisted peer review», Humanit. Soc. Sci. Commun., vol. 8, n.o 1, p. 25, ene. 2021, doi: 10.1057/s41599-020-00703-8.

[35] S. Azad et al., «A Reviewer Recommender System for Scientific Articles Using a New Similarity Threshold Discovery Technique», en Proceedings of the Fourth International Conference on Trends in Computational and Cognitive Engineering, Springer, Singapore, 2023, pp. 503-518. doi: 10.1007/978-981-19-9483-8_42.

[36] J. Devlin, M.-W. Chang, K. Lee, y K. Toutanova, «BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding», presentado en Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), jun. 2019, pp. 4171-4186. doi: 10.18653/v1/N19-1423.

[37] M. Kubát y S. Matwin, «Addressing the Curse of Imbalanced Training Sets: One-Sided Selection», presentado en International Conference on Machine Learning, 1997. Accedido: 21 de noviembre de 2025. [En línea]. Disponible en: https://www.semanticscholar.org/paper/Addressing-the-Curse-of-Imbalanced-Training-Sets%3A-Kub%C3%A1t-Matwin/ebc3914181d76c817f0e35f788b7c4c0f80abb07

[38] N. V. Chawla, K. W. Bowyer, L. O. Hall, y W. P. Kegelmeyer, «SMOTE: Synthetic Minority Over-sampling Technique», J. Artif. Intell. Res., vol. 16, pp. 321-357, jun. 2002, doi: 10.1613/jair.953.

[39] T. Wongvorachan, S. He, y O. Bulut, «A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining», Information, vol. 14, n.o 1, p. 54, ene. 2023, doi: 10.3390/info14010054.

[40] L. Xu, M. Skoularidou, A. Cuesta-Infante, y K. Veeramachaneni, «Modeling Tabular data using Conditional GAN», 28 de octubre de 2019, arXiv: arXiv:1907.00503. doi: 10.48550/arXiv.1907.00503.

[41] M. E. S.-G. González-Pérez Pedro P., «Addressing the class imbalance in tabular datasets from a generative adversarial network approach in supervised machine learning - Máximo E Sánchez-Gutiérrez, Pedro P González-Pérez, 2023», J. Algorithms Comput. Technol., nov. 2023, Accedido: 5 de mayo de 2026. [En línea]. Disponible en: https://journals.sagepub.com/doi/10.1177/17483026231215186

[42] J. M. Johnson y T. M. Khoshgoftaar, «Survey on deep learning with class imbalance», J. Big Data, vol. 6, n.o 1, p. 27, mar. 2019, doi: 10.1186/s40537-019-0192-5.

[43] G. Kovács, «An empirical comparison and evaluation of minority oversampling techniques on a large number of imbalanced datasets», Appl. Soft Comput., vol. 83, p. 105662, oct. 2019, doi: 10.1016/j.asoc.2019.105662.

[44] A. C. Ribeiro, A. Sizo, H. Lopes Cardoso, y L. P. Reis, «Acceptance Decision Prediction in Peer-Review Through Sentiment Analysis», en Progress in Artificial Intelligence, Springer, Cham, 2021, pp. 766-777. doi: 10.1007/978-3-030-86230-5_60.

[45] P. K. Bharti, T. Ghosal, M. Agrawal, y A. Ekbal, «How Confident Was Your Reviewer? Estimating Reviewer Confidence from Peer Review Texts», en Document Analysis Systems, Springer, Cham, 2022, pp. 126-139. doi: 10.1007/978-3-031-06555-2_9.

[46] A. Kumar, T. Ghosal, y A. Ekbal, «A Deep Neural Architecture for Decision-Aware Meta-Review Generation», en 2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL), sep. 2021, pp. 222-225. doi: 10.1109/JCDL52503.2021.00064.

[47] A. Kumar, T. Ghosal, S. Bhattacharjee, y A. Ekbal, «Towards automated meta-review generation via an NLP/ML pipeline in different stages of the scholarly peer review process», Int. J. Digit. Libr., vol. 25, n.o 3, pp. 493-504, sep. 2024, doi: 10.1007/s00799-023-00359-0.

[48] P. K. Bharti, M. Navlakha, M. Agarwal, y A. Ekbal, «PolitePEER: does peer review hurt? A dataset to gauge politeness intensity in the peer reviews», Lang. Resour. Eval., vol. 58, n.o 4, pp. 1291-1313, dic. 2024, doi: 10.1007/s10579-023-09662-3.

[49] A. Kumar, T. Ghosal, S. Bhattacharjee, y A. Ekbal, «Investigations on Meta Review Generation from Peer Review Texts Leveraging Relevant Sub-tasks in the Peer Review Pipeline», Lect. Notes Comput. Sci. Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinforma., vol. 13541 LNCS, pp. 216-229, 2022, doi: 10.1007/978-3-031-16802-4_17.

[50] C.-H. Lai y P.-Y. Peng, «A Hybrid Deep Learning Method to Extract Multi-features from Reviews and User–Item Relations for Rating Prediction», Int. J. Comput. Intell. Syst., vol. 16, n.o 1, p. 109, jun. 2023, doi: 10.1007/s44196-023-00288-5.

[51] V. Lakshmanan, S. Robinson, y M. Munn, «Machine Learning Design Patterns».

[52] «Microsoft Word - ANEXO 7. AREAS OCDE». Accedido: 21 de noviembre de 2025. [En línea]. Disponible en: https://minciencias.gov.co/sites/default/files/upload/convocatoria/anexo_7._areas_ocde_0.pdf

[53] «Convocatoria de Clasificación y Reconocimiento de Revistas Científicas Nacionales - Publindex 2026», Minciencias. Accedido: 8 de julio de 2026. [En línea]. Disponible en: https://minciencias.gov.co/convocatorias/convocatoria-clasificacion-y-reconocimiento-revistas-cientificas-nacionales-publindex

[54] H. He y E. A. Garcia, «Learning from Imbalanced Data», IEEE Trans. Knowl. Data Eng., vol. 21, n.o 9, pp. 1263-1284, sep. 2009, doi: 10.1109/TKDE.2008.239.

[55] R. Kohavi, «A study of cross-validation and bootstrap for accuracy estimation and model selection», en Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2, en IJCAI'95. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., ago. 1995, pp. 1137-1143.

[56] J. Dodge, G. Ilharco, R. Schwartz, A. Farhadi, H. Hajishirzi, y N. Smith, «Fine-Tuning Pretrained Language Models: Weight Initializations, Data Orders, and Early Stopping», arXiv.org. Accedido: 7 de mayo de 2026. [En línea]. Disponible en: https://arxiv.org/abs/2002.06305v1

[57] M. Sokolova y G. Lapalme, «A systematic analysis of performance measures for classification tasks», Inf. Process. Manag., vol. 45, n.o 4, pp. 427-437, jul. 2009, doi: 10.1016/j.ipm.2009.03.002.

[58] G. C. Cawley y N. L. C. Talbot, «On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation», J Mach Learn Res, vol. 11, pp. 2079-2107, ago. 2010.

[59] J. T. Hancock y T. M. Khoshgoftaar, «CatBoost for big data: an interdisciplinary review», J. Big Data, vol. 7, n.o 1, p. 94, nov. 2020, doi: 10.1186/s40537-020-00369-8.

[60] L. Yang y A. Shami, «On Hyperparameter Optimization of Machine Learning Algorithms: Theory and Practice», arXiv.org. Accedido: 7 de mayo de 2026. [En línea]. Disponible en: https://arxiv.org/abs/2007.15745v3

[61] L. Breiman, «Random Forests», Mach. Learn., vol. 45, n.o 1, pp. 5-32, oct. 2001, doi: 10.1023/A:1010933404324.

[62] R. Gopalan, D. Onniyil, G. Viswanathan, y G. Samdani, «Hybrid models combining explainable AI and traditional machine learning: A review of methods and applications», World J. Adv. Eng. Technol. Sci., vol. 15, pp. 1388-1402, may 2025, doi: 10.30574/wjaets.2025.15.2.0635.

[63] R. Blagus y L. Lusa, «SMOTE for high-dimensional class-imbalanced data», BMC Bioinformatics, vol. 14, n.o 1, p. 106, mar. 2013, doi: 10.1186/1471-2105-14-106.

[64] D. H. Wolpert, «La falta de distinciones a priori entre algoritmos de aprendizaje | Computación neuronal | MIT Press», vol. 8, n.o 7, Neural Computation, 1 de octubre de 1996. Accedido: 8 de mayo de 2026. [En línea]. Disponible en: https://direct.mit.edu/neco/article-abstract/8/7/1341/6016/The-Lack-of-A-Priori-Distinctions-Between-Learning?redirectedFrom=fulltext

DECLARATION OF ARTIFICIAL INTELLIGENCE USE

The AI tool (CHATGPT) was used only to guide the design of graphic material (Fig.1 and Fig 2); this instrument did not participate in the writing of the text, in the data analysis, in the programming of the models, or in the interpretation of the results presented in this manuscript, which were prepared entirely by the authors.

ACKNOWLEDGMENTS

The authors thank the journal editor who took part in this project, whose information was the central input. This work was carried out as a thesis project to obtain the Master's degree in Artificial Intelligence from Universidad de la Salle.

FUNDING

No financing

CONFLICT OF INTEREST

None

AUTHORSHIP CONTRIBUTIONS

Drafting – original draft: Liseth Paola Claro Ascanio, Cristian Camilo Tirado Cifuentes.

Writing–review and editing: Liseth Paola Claro Ascanio, Cristian Camilo Tirado Cifuentes.