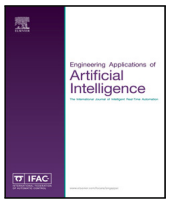
Engineering Applications of Artificial Intelligence 156 (2025) 111039

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai



Research paper

Automated research methodology classification using machine learning☆

Zsolt T. Kosztyán [a,b,*,1], Tünde Király [a,2], Tibor Csizmadia [c,3], Attila Imre Katona [a,4], Ágnes Vathy-Fogarassy [d,5]

[a] Department of Quantitative Methods Faculty of Business and Economics University of Pannonia, Egyetem str. 10., Veszprém, Hungary
[b] Institute of Advanced Studies, Kőszeg (iASK), Chernel str. 10, 9730, Kőszeg, Hungary
[c] Department of Management, Institute of Management Faculty of Business and Economics University of Pannonia, Egyetem str. 10., Veszprém, Hungary
[d] Department of Computer Science Faculty of Information Technology University of Pannonia, Egyetem str. 10., Veszprém, Hungary

ARTICLE INFO

Keywords:
Machine learning
Classification
Applied research methods
Extreme Gradient Boosting
Large language models

ABSTRACT

Scientific papers have become the primary means for disseminating scientific research, and thus, the ability to classify research papers based on different aspects has become essential. Therefore, many works have developed classification approaches; however, they focused solely on research topic-based classification. In addition, no solution has been developed to classify papers based on the applied methodology, and finally, the accuracy of the existing paper classification methods is not satisfactory. In this study, a novel automated classification methodology using a refined Extreme Gradient boosting (XGBoost) model is presented to classify the research methods employed in scientific papers. Three article sets, including quantitative and qualitative research methods, were collected from the topics of tourism, medical science and information systems, consisting of 229, 557 and 787 papers, respectively. The classification problem was considered a binary classification task to maintain interpretability. The developed model was trained and tested on article set 1 (tourism) and 2 (medical science), and then, the proposed model was applied to article set 3, (information systems and tourism). The high accuracy achieved in different research fields (90%–95% accuracies on average) indicates that the proposed classification model is generalizable because it can be successfully applied in many disciplines. The automated classifier enables the rapid acquisition of vital information and the identification of significant differences among the applied methodologies in various research domains. A future development direction will be to increase the scalability of the proposed model to achieve efficient operations on large volumes of research papers.

1. Introduction

Academic papers have become the primary means for disseminating scientific research (Harzing and Adler, 2016; Zhang et al., 2021). The large number of research papers published in the last decade highlights the need for a quick and reliable overview of the field, which is fundamental for preparing a valid and novel academic study (Atkinson, 2024; Zhang et al., 2022). Moreover, researchers must spend considerable time finding appropriate publications that support their research (Chowdhury and Schoen, 2020; Wolfswinkel et al., 2013). Thus, automatic research paper classification has become increasingly important (Shu et al., 2020), and the classification of research papers on the basis of different characteristics is essential in valid academic analyses and research assessments (Zhang et al., 2022) and effectively supports developing research fields (Rivest et al., 2021).

The process of manually classifying research papers is extremely time-consuming and often involves unnecessary human errors (Chowdhury and Schoen, 2020). Furthermore, the long processing time required for manual classification leads to significant delays when disseminating valuable scientific results (Harzing and Adler, 2016). Therefore, an automated method for processing, assessing and analyzing

☆ This document presents the results of the research project funded by the Hungarian National Science Foundation (OTKA-143482).

* Corresponding author at: Institute of Advanced Studies, Kőszeg (iASK), Chernel str. 10, 9730, Kőszeg, Hungary.
E-mail addresses: kosztyan.zsolt@gtk.uni-pannon.hu (Z.T. Kosztyán), kiraly.tunde@gtk.uni-pannon.hu (T. Király), csizmadia.tibor@gtk.uni-pannon.hu (T. Csizmadia), katona.attila@gtk.uni-pannon.hu (A.I. Katona), vathy.agnes@gtk.uni-pannon.hu (Á. Vathy-Fogarassy).
[1] Full Professor, Research Fellow.
[2] PhD. student, Research assistant.
[3] Head of Institute, Professor.
[4] Head of Department, Senior Research Fellow.
[5] Head of Department, Associate Professor.

https://doi.org/10.1016/j.engappai.2025.111039
Received 8 September 2024; Received in revised form 8 April 2025; Accepted 28 April 2025
Available online 31 May 2025

many research papers so that they can be quickly and accurately classified is critical (Kim and Gil, 2019).

Automatic research paper classification is performed via unsupervised and supervised methods (Mendoza et al., 2022). Some researchers have attempted to categorize research papers with unsupervised methods, for example, by categorizing them through semantic analyses (Hanyurwimfura et al., 2014; Salatino et al., 2022), keyword co-occurrence structures (Kim and Gil, 2019), citation relations (Van Eck and Waltman, 2017), or word embedding-based community detection techniques (Gündoğan and Kaya, 2020). The algorithms used in these studies include k-means, Louvain community detection, the Girvan–Newman method, and latent Dirichlet allocation (LDA).

Numerous studies have aimed to classify research papers on the basis of their topics via supervised learning methods. For example, Taheriyan (2011) proposed a graph-based supervised classification method to measure the similarity between research papers and to categorize them on the basis of their subjects. Waltman and Van Eck (2012) proposed a subject area-based classification method for research papers by using the information obtained from their citation structures. Eykens et al. (2021) developed a naïve Bayes multilabel classifier to assign subdisciplines to research papers on the basis of their abstracts and titles. Although these studies presented promising results, the accuracies achieved by these models were not satisfactory, as stated by Daradkeh et al. (2022).

To overcome this issue, more accurate paper classification models have been developed using Deep Learning (DL) methods that classify research papers based on their topics. For example, Rivest et al. (2021) and Daradkeh et al. (2022) proposed convolutional neural network (CNN)- based paper classification methods, whereas Hoppe et al. (2021) suggested the use of recurrent neural network (RNN) models operating on knowledge graphs to classify scientific works. Similarly, Kandimalla et al. (2021) applied a bidirectional RNNs model; however, in their solution, an attention layer followed the two preceding RNNs layers. Transformer-based methods have also been used because of their fast calculation times and outstanding classification performance (Al Fahoum, 2023); however, the related studies focused mainly on automated academic writing (Golan et al., 2023; Hutson, 2022; de la Torre-López et al., 2023) and news classification (Shushkevich et al., 2023). In the realm of transformer-based models, Bidirectional Encoder Representations from Transformers (BERT) has proven to be more effective for classification tasks than Generative Pre-trained Transformers (GPT) (Lukito et al., 2024; De Santis et al., 2025). This is due mainly to the bidirectional nature of BERT, which allows one to fully understand the contextual relationships between the words in a given text. This has resulted in BERT-based solutions being commonly chosen for scientific article classification tasks, as GPT models, owing to their structures, excel more in text generation and one-shot or few-shot learning tasks (De Santis et al., 2025). For example, models based on BERT were used to improve the performance achieved in a literature review task (Khadhraoui et al., 2022), in a research field classification study (Wolff et al., 2024), and in an analysis of categories and trends (Raja et al., 2024). One limitation of transformer-based models in scientific article classification applications is their dependency on large, well-annotated datasets, which can be challenging to obtain in niche research domains, potentially leading to biased or inaccurate classifications in underrepresented fields. On the other hand, transformer-based models, such as BERT and GPT, offer excellent scalability for large-scale text processing tasks but require substantial computational resources, including high memory capacities and the use of specialized hardware (such as graphics processors or tensor processing units) for acceleration, making their deployment and fine-tuning steps computationally expensive. Table 1 summarizes the studies that have addressed research paper classification.

All the previously mentioned studies focused on the classification of research papers based on their fields or topics; however, no research paper classification method that categorizes papers based on other features, such as their research methodologies, has been developed. In addition, during a scientometric analysis, the classification process is only partially supported or automated since the support can be utilized only for classifying papers according to their topics or fields. Finally, in prior studies, only the titles, keywords, or abstracts of the examined papers were considered, and no works have investigated whether the classification performance could be improved if the entire text were preprocessed and considered (see Table 1). Furthermore, as Table 1 shows, during the evaluation phase, either one specific research topic was selected or the papers were collected initially from mixed topics; however, no study has analyzed how the proposed methods performed in different research fields, i.e., the generalizability of the proposed classification methods.

On the basis of the challenges described in the state-of-the-art literature, we synthesized Fig. 1 to present the main problem addressed in this paper.

Fig. 1 shows the context of the research gaps at three levels: (1) the main cause, (2) the resulting problems, and (3) the consequences of these problems. The cause of these problems is that no comprehensive automatic classification models are available since the existing models focus only on classifying papers by their fields or research topics (cause $(CA)_1$). Notably, research topic-based classification is important; however, doing so does not provide information about the other features of the examined papers, such as the research methodologies used. Nevertheless, important features such as the applied research methodologies are extremely important since changes, trends, or milestones in terms of how the studies were conducted cannot be determined without these features. However, scientometric analyses focus solely on topic classification, and the research method applied in each paper must be identified manually due to the lack of comprehensive classification models, which is a very time-consuming process (problem $(PR)_1$). Furthermore, manually classifying features other than research topics leads to the problem that only a limited number of papers can be processed due to human work capacity limitations (PR_2). The results of manual classification are subjective since they are distorted by the decision maker's knowledge and intuitions, weakening the reproducibility of these results (PR_3). In addition, the existing models have only been tested in specific research fields, or test data were collected initially from mixed research fields. The previous studies did not analyze how these models performed depending on the context of the corpus (the research field of the papers to be classified); therefore, their generalizability was not guaranteed (PR_4). Finally, the existing methods do not use the full body of each paper, which leads to information losses (PR_5).

As a consequence of the aforementioned problems, researchers cannot perform comprehensive scientometric analyses on large datasets since only topic classification is supported by the existing approaches (consequence $(CO)_1$). Although comprehensive scientometric analysis techniques have been developed, the dissemination of such techniques takes a long time because of delays caused by human work capacity limitations (CO_2). Furthermore, owing to the lack of an automatic research method classifier, the current scientometric analyses are not extensive because they do not focus on both research topic and methods (CO_3). Finally, subjective human judgments can lead to distorted or not completely valid results, weakening the reproducibility of the classification results (CO_4).

The goal of our research is to examine whether the automatic classification of scientific articles on the basis of their research methodologies can be achieved via classic ML algorithms. The reason for preferring classic machine learning algorithms over transformer models is simple. Transformer models require a large amount of training data to accurately capture the contextual relationships within articles, and this also comes with significant computational demands. Our proposed research goal focuses on the methodological classification of the research methods used in scientific articles, which does not require comprehensive processing and understanding of each entire article. Instead, it is sufficient to extract only the relevant details that pertain to

Table 1
Studies that have addressed scientific paper classification.

Type[a]	Reference	Method[b]	Data[c]	Topic	Domain[d]					
					TI	KW	AB	AU	CI	FT
S	Taheriyan (2011)	Relationship graphs	ACM	mixed	✓			✓		
	Waltman and Van Eck (2012)	Hierarchical classification	WoS	astronomy					✓	
	Eykens et al. (2021)	Naïve Bayes	ProQuest	sociology	✓		✓			
	Rivest et al. (2021)	CNN	Scopus	mixed	✓	✓	✓	✓		
	Daradkeh et al. (2022)	CNN	Scopus, ProQuest, EBSCOhost	mixed	✓	✓	✓	✓		
	Hoppe et al. (2021)	RNN	ArXiv	computer vision, mathematics	✓		✓	✓		
	Kandimalla et al. (2021)	RNN	CiteSeerX	mixed			✓			
	Khadhraoui et al. (2022)	BERT	PubMed	virology		✓				
	Wolff et al. (2024)	BERT	ORKG, ArXiv, Semantic Scholar	mixed	✓		✓			
	Raja et al. (2024)	BERT	PubMed, WoS	ocular diseases			✓			
U	Hanyurwimfura et al. (2014)	K-means	SCIRP publisher	computer science	✓	✓				
	Yohan et al. (2014)	NER	TeluguWiki	mixed		✓				
	Salatino et al. (2022)	CSO	MAG	data mining	✓	✓	✓	✓		
	Van Eck and Waltman (2017)	Community detection	WoS	astronomy					✓	
	Kim and Gil (2019)	LDA	FGCS journal	computer science	✓	✓	✓			
	Gündoğan and Kaya (2020)	Community detection	ArXiv	mixed	✓		✓			

[a] *S=Supervised, U=Unsupervised.*
[b] *NER=Named Entity Recognition, CSO=Computer Science Ontology.*
[c] *ACM=Association for Computing Machinery, WoS=Web of Science, EBSCOhost=Elton B. Stephens Company dataset, ORKG=Open Research Knowledge Graph, PubMed=Public Medical, SCIRP=Scientific Research Publishing, TeluguWiki = Telugu Wikipedia, MAG=Microsoft Academic Graph, FGCS=Future Generation Computer Systems.*
[d] *TI=Title, KW=Keywords, AB=Abstract, AU=Authors, CI=Citation, FT=Full text.*

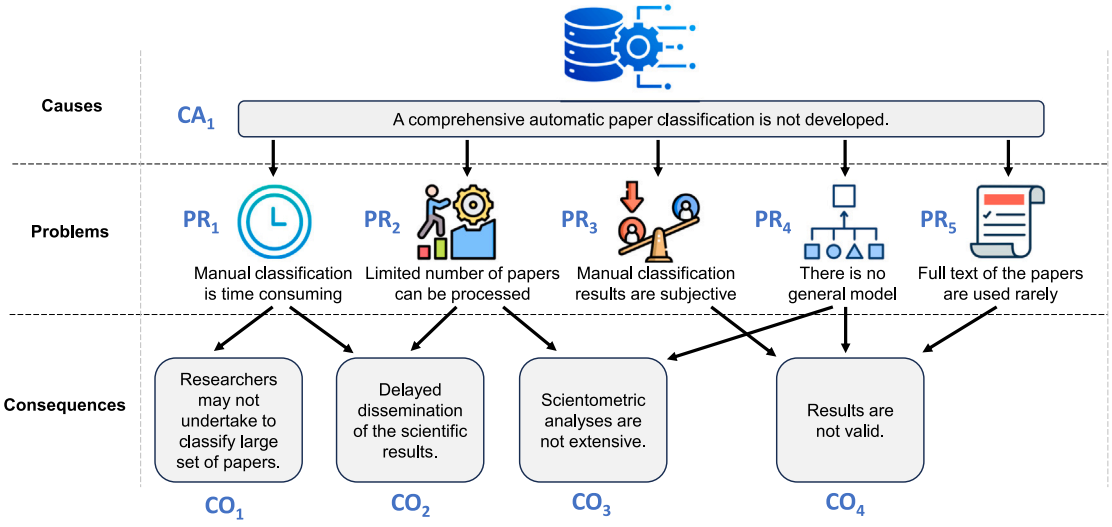


Fig. 1. Problem description.

the applied methodology. These objectives can likely be achieved using simpler machine learning models, which require less training data and significantly lower computational resources than transformer models do. In this paper, we consider the problem to be a binary classification problem in which each research methodology is quantitative or qualitative. There are two main reasons for this decision. First, by concentrating exclusively on qualitative and quantitative research, we can delineate distinct categories that enhance the interpretation and analysis processes. This clarity is essential, particularly for extensive analyses where subtle classifications may result in ambiguity. Second, when classification problems are structured to differentiate between only two categories, machine learning models frequently exhibit superior performance due to the decreased complexity of the resulting feature space. The complexities and subtleties of mixed-method classification might hinder the learning process; thus, models can be more efficiently optimized when focused on differentiating between two methodologies. In our work on developing an article classification model based on the applied research methodologies, we investigate the following research questions (RQs).

RQ_1 Can the process of classifying the research methods employed in articles be automated via classic machine learning methods? If so, which classic machine learning algorithm is most efficient for classifying the research methodologies used in scientific articles, and what level of accuracy can be achieved?

RQ_2 Can article classification be performed solely on the basis of abstracts, or is it necessary to process the entire text of the articles for classification purposes? Furthermore, what text preprocessing steps should be performed before conducting classification?

RQ_3 Should corpus-dependent models be created for specific research fields, or is the developed classification model generalizable?

The associated contributions (Cs) of the paper are as follows.

C_1 We propose an automated research method classification model that supports scientometric analyses.

C_2 We compare the classification performance of abstract-based and full-text-based models and determine the text preprocessing tasks that are necessary prior to applying the developed classification model.

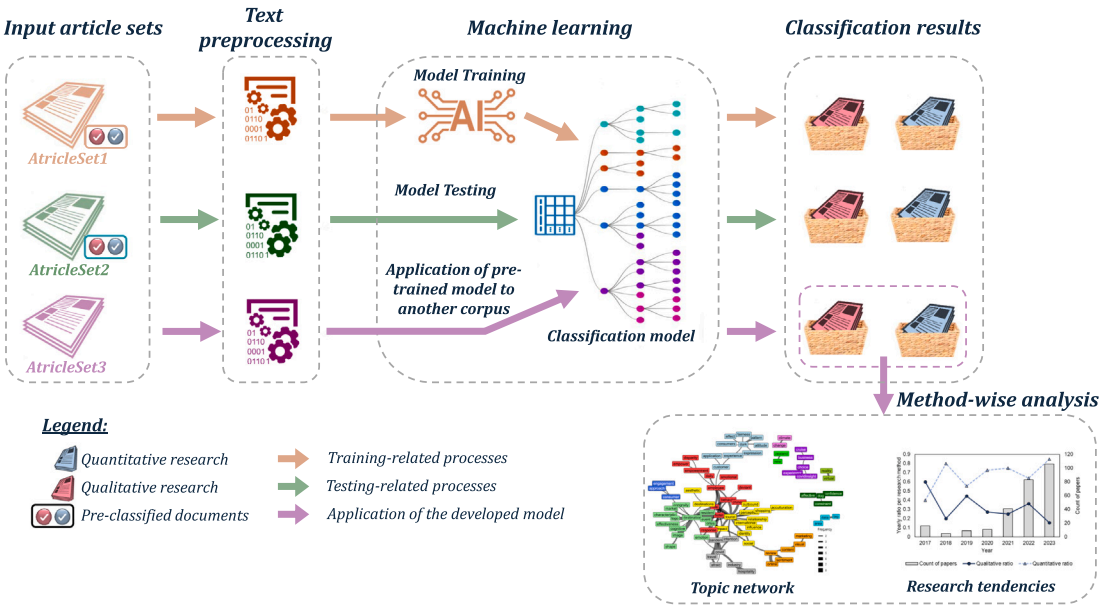


Fig. 2. Research framework.

C_3 We investigate whether the proposed model can be generalized across different fields.

The development of an automated research method classification model (C_1) is beneficial from various perspectives. Scholars can classify large sets of papers, the dissemination time required for new approaches decreases because less manual work is needed, and more comprehensive analyses can be performed while also considering the research methodologies of the papers, not just their topics. Identifying which sections of the papers should be used for classification and determining the appropriate preprocessing methods (C_2) facilitates the development of more effective classification models, thereby enhancing the validity of the obtained scientometric analysis results. Moreover, determining whether the proposed model is generalizable across different fields is important in terms of model development (C_3) since, with this knowledge, research classification models can be better designed in diverse research fields.

Compared with the existing models, the proposed model grants new opportunities since it enables the classification of papers on the basis of not only their research methods but also their topics, considers the entire content of each paper and not only its abstract and/or title, is able to achieve higher classification accuracy and is highly generalizable.

The remainder of this paper is organized as follows. The research method is introduced in Section 2. Section 3 presents the results of the study, and Section 4 presents a comprehensive case study as an example of the proposed methodology and model. In Section 5, the results of the study are discussed, and finally, in Section 6, a summary and the conclusions of the paper are presented.

2. Methods

2.1. Research framework

For clarity, we introduce the research framework first, as shown in Fig. 2.

To build an effective classification methodology and machine learning model, three article sets were used as input datasets (detailed descriptions are provided in Section 2.2). The first two article sets were used for training, validating, and testing the classification model to be developed. The articles contained in these article sets had been previously categorized into two classes on the basis of their applied research methodologies: articles presenting quantitative or qualitative research. In the third article set, the papers were not prelabeled on the basis of their research methodologies.

To construct an effective classification model, different machine learning algorithms were fine-tuned and tested. The performance of different classification models was compared by conducting a quantitative analysis on the classification results obtained from the first and second article sets (a detailed description is provided in Section 2.3). To demonstrate the applicability of the proposed model, Article Set 3 was classified using the developed model, and the classified papers were further analyzed to identify the research tendencies and key topics within each predicted class.

2.2. Data employed

In this research, three sets of articles were applied, and the goal was to classify the included papers on the basis of the utilized methodology, i.e., quantitative or qualitative research. The first set of articles (*Article Set 1*) was used to develop the classification process, including determining the necessary text preprocessing steps, fine-tuning and comparing the effectiveness of different machine learning algorithms, and testing the resulting machine learning models. As a result, we established a methodology for classifying scientific articles and developed a classifier that was effective at classifying the articles contained in the first set.

To answer the third research question (RQ_3), a second set of articles (*Article Set 2*) was also considered. These articles were used to test the effectiveness of the developed methodology and machine learning models in terms of classifying articles with different subjects. The success achieved when classifying this set of articles indicated the universal applicability of the developed methodology and models.

The first two sets of articles were used to develop, validate, and evaluate the usability of the model. To achieve these goals, it was necessary to know the classifications (class labels) of the articles. As the model was developed to classify scientific articles according to the research methodology used, we first determined whether a qualitative or quantitative methodology was applied in each study. This classification step was performed manually by individually examining each

Table 2
Main characteristics of the article sets used in this research. The N/R designation indicates values that were not relevant to the study.

Characteristic	*Article set 1*	*Article set 2*	*Article set 3*
Number of papers	229	557	787
Number of papers using quantitative research methods	149	546	N/R
Number of papers using qualitative research methods	80	11	N/R
Median article length (number of words)	9,404	5,970	12,297
Median article lengths (number of characters)	63,694	39,718	85,368
Length of the shortest article (number of words)	2,583	124	77
Length of the shortest article (number of characters)	17,993	934	605
Length of the longest article (number of words)	30,519	73,823	57,671
Length of the longest article (number of characters)	213,481	513,739	399,433
Number of papers with abstracts	223	N/R	N/R
Number of abstracts in papers using quantitative research methods	145	N/R	N/R
Number of abstracts in papers using qualitative research methods	78	N/R	N/R
Median abstract length (number of words)	187	N/R	N/R
Median abstract length (number of characters)	1,298	N/R	N/R

article. This information served as class labels for the machine learning algorithms during the training, validation, and testing phases.

The third set of articles (*Article Set 3*) differs from the previous two sets in several ways. First, the topic of this set of articles differs from the topics of the first two sets. Second, no manual classification process was carried out on these articles. We used the third group of articles to provide examples of how researchers can apply the methodologies and models developed in their work. After the classification task was performed, 100 randomly chosen articles were manually checked.

To test the generalizability of the proposed method, the three groups of articles were obtained from different scientific fields. The first group of articles collected by Sulyok et al. (2023) contains articles that were published in the field of tourism concerning coronavirus disease (COVID)-19 and classified on the basis of their research methodologies and other factors, such as the data source, target group, and focus/destination of each article. The second group of articles, which was also manually classified by Churruca et al. (2021), contains articles that were published primarily in health and medical science. These articles focus on safety in hospitals. The third group of articles contains all the articles published by the International Journal of Information Management and Tourism Management from 2017–2023. The main characteristics of the three article sets are summarized in Table 2. It includes details such as the number of papers, the research methodology (quantitative or qualitative), the median article and abstract lengths (in words and characters), and the shortest and longest articles in each set.

The Article Sets are freely accessible on the Figshare platform (Katona et al., 2025).

2.3. Methods employed

In this work, a binary classification methodology was developed, and a classification model was designed and tested. The goal of the developed methodology and model was to distinguish whether the published articles employ a quantitative or qualitative research methodology. The articles employing quantitative research methods were considered the negative class and were assigned the label "0", whereas the articles utilizing qualitative methods were classified as the positive class and were assigned the label "1".

2.3.1. Investigated machine learning algorithms

To address the first question (RQ_1), we tested and evaluated the following machine learning algorithms: Logistic regression (LR), Support vector machine (SVM), Random forest (RF), Extreme Gradient Boosting (XGBoost), and artifical neural network (ANN).

Logistic regression is a statistical method that is used mainly for binary classification. This method can be used to predict the probability of an observation belonging to a particular category. The relationships between the independent variables and the binary outcome are modeled by using the logistic function, ensuring that the predictions fall within the range of 0–1, which represents the probability of belonging to the positive class. During the learning process, the coefficients of the logistic function are estimated by optimizing the fit between the expected outputs (classifications) and the predicted probabilities of belonging to a particular class.

Support vector machines are also statistical machine learning algorithms. The aim of the SVM algorithm is to find a hyperplane in a high-dimensional space that best separates the given data points into different classes. The optimal decision boundary (hyperplane) is identified by selecting support vectors that include the data points that are closest to the decision boundary. An SVM is particularly efficient in terms of handling high-dimensional data, which is one of the main features of our input dataset.

The random forest and *XGBoost* are both ensemble learning algorithms that are based on decision trees and utilize tree structures for prediction purposes. The RF model is an ensemble method that combines multiple decision trees (weak learners) and uses their average or a majority vote to determine the output class or value. During the process of constructing a classification model (a forest of decision trees), the decision trees are built independently, and random sampling is applied to the training samples and input features. XGBoost is also an ensemble method that uses decision trees for classification, but in contrast with the RF algorithm, XGBoost is a boosting algorithm that aims to correct the errors induced by the previous models.

Artificial neural networks are machine learning models inspired by the structure and functioning of the human brain. They consist of interconnected nodes (neurons) that are organized in layers. The learning algorithm of a neural networks aim to learn the strengths of the connections between the neurons and the bias values of the neurons. Neural networks are capable of learning intricate patterns and relationships in data, making them suitable for complex classification tasks. However, these methods are black-box models, making it challenging to interpret them.

2.3.2. Classification model development methodology

The same approach for preprocessing the articles was used for all the machine learning algorithms. The articles were converted from their original PDF formats to plain text files. According to research question RQ_2, when only the abstracts were utilized for the classification process, only the text of the abstracts was converted to plain text; otherwise, the whole content of each article was converted to plain text. To develop the most appropriate classification methodology, several additional text preprocessing steps were also performed. First, the text was converted to lowercase. As an optional text preprocessing step aimed at improving the accuracy of the classification model, we also performed text lemmatization and analyzed its effect on the resulting classification accuracy. Lemmatization is the process of transforming text by breaking words down into their root meanings. Examining the impact of lemmatization is crucial for the development of a classifier model for scientific articles, as it helps normalize word variations, reduce vocabulary sizes, and improve the ability of the

model to recognize key terms and conceptual relationships, ultimately enhancing the attained classification accuracy. The next preprocessing step in all cases involved tokenizing the transformed text, which, in this case, meant breaking the text into words. As the last preprocessing step, a document-term matrix was created to generate the appropriate structured input for the machine learning algorithms. Each row in the matrix represented one article, and the columns defined predetermined terms (words or sequences of words). The matrix values represented the absolute frequencies of the given terms occurring in the respective documents. The terms used in the columns of the document-term matrix were manually defined by collecting 307 terms that were considered representative in the classification task. The list of terms was created by merging the following glossaries. We used the Colorado State University glossary file of qualitative and quantitative research (Colorado State University, 2024). This glossary provides definitions of many of the terms used in guides for conducting qualitative and quantitative research. We also included Tennessee Tech University's vocabulary (Tennessee Tech University, 2024) and the Content Authority lists of words for quantitative (The Content Authority, 2024) and qualitative research (The Content Authority, 2024). We merged these glossaries by excluding duplicates. In this way, we obtained 307 terms related to qualitative and quantitative research. The list of terms is provided in Appendix A. Notably, to ensure correct processing, the word list presented in Appendix A was also converted to lowercase when the document-term matrix was created.

To develop classification models, *Article Set 1* was randomly divided into training and test sets stratifying the class label attributes. Eighty percent of the available articles were used for training the algorithms, and the remaining 20% were used for testing. To address the imbalance in *Article Set 1*, random oversampling was applied to the training set to adjust the ratio between the minority and majority classes.

During the model development procedure, all machine learning algorithms were fine-tuned if needed. As the logistic regression algorithm did not require hyperparameter tuning, hyperparameter tuning was not performed for this algorithm. In addition, for this algorithm, the maximum number of iterations for the learning process was set to 1000. The SVM method was tested only as a linear classification method, so no kernel was applied. To tune the RF and XGBoost algorithms and find the most appropriate neural network model, the hyperparameters presented in Table 3 were fine-tuned, along with their respective search spaces, which are also displayed in the table. The fine-tuning step was carried out by performing a grid search on the search spaces and optimizing the area under the receiver operating characteristic curve (AUC) values. The neural network was trained for a maximum of 50 epochs. To prevent overfitting, early stopping regularization with a patience level of 5 epochs was applied.

The complete workflow of the article classification model development process is summarized in Fig. 3.

2.3.3. Evaluation methods for the developed classification models

During the model development phase, the performance of the classification models was evaluated on test sets derived from *Article Set 1*. To reduce the bias induced by randomly generated test sets, the efficiency of the machine learning algorithms was determined on the basis of multiple evaluations. Specifically, for each machine learning algorithm, we performed 1000 random splits to divide the article set into training and test datasets. Across these 1000 splits, 1000 models were created, and the performance of each model was evaluated on the basis of the test set. The final evaluation metrics were determined as the averages calculated from these 1000 evaluations.

The model evaluations were performed using the following quality measures: accuracy, sensitivity (recall), specificity, the F1 score, the average F1 score, and the AUC. The values of the first four evaluation metrics were computed as follows (Tasci et al., 2024):

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

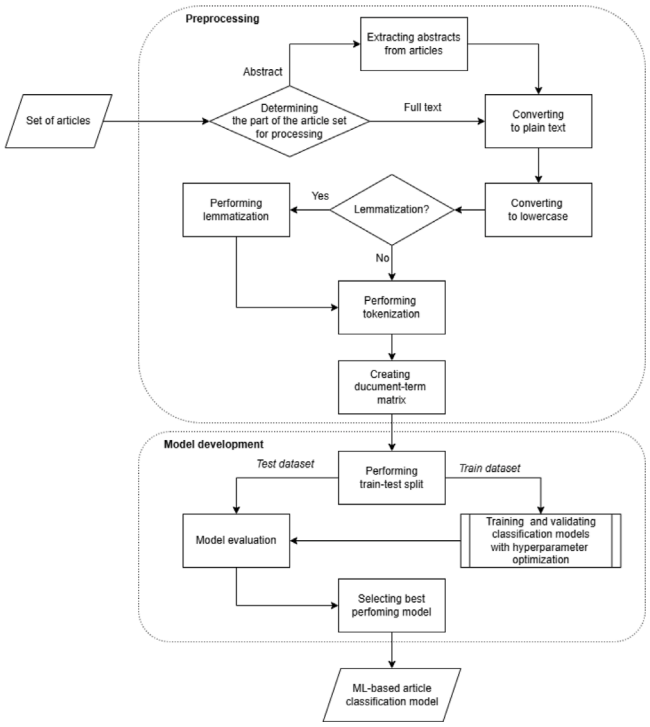


Fig. 3. Workflow of the development process of the article classification model.

$$sensitivity = \frac{TP}{TP + FN} \tag{2}$$

$$specificity = \frac{TP}{TP + FN} \tag{3}$$

$$F1 = \frac{2TP}{2TP + FP + FN} \tag{4}$$

where true positive (TP) represents the number of samples that were correctly classified as belonging to the positive class (qualitative class), true negative (TN) represents the number of samples that were correctly classified as belonging to the negative class (quantitative class), false positive (FP) represents the number of samples belonging to the negative class that were classified by the model as belonging to the positive class, and false negative (FN) represents the number of samples belonging to the positive class that were classified by the model as belonging to the negative class.

When performing binary classification with a heavily imbalanced test set, the traditional F1 score might not adequately reflect the performance of the tested model, as this metric tends to be overly influenced by the majority class. In contrast, to compute the average F1 score, F1 scores are calculated separately for each class, and then these scores are averaged on the basis of the frequencies of the classes; thus, the average F1 score enables a more balanced evaluation. This is particularly advantageous in scenarios with imbalanced data, as this metric is useful for evaluating the effectiveness of a model across both positive and negative classes, offering a more comprehensive measure of its classification performance.

AUC represents the overall performance of the model in terms of distinguishing between positive and negative classes across different thresholds. In the receiver operating characteristic curve (ROC) curve, the true-positive rate (sensitivity) was plotted against the false-positive rate, illustrating the tradeoff between sensitivity and specificity. The AUC is the area under the ROC curve, and this value ranges from 0–1. In an ideal classification scenario, the AUC value would be 1, indicating perfect discrimination between positive and negative instances. The

Table 3
The tuned hyperparameters of the machine learning algorithms and the search ranges applied during the tuning process.

Algorithm	Hyperparameter	Values
RF	Number of estimators	20, 30, 40, 50, 75, 100, 120
	Splitting criterion	gini, entropy
	Max tree depth	10, 12, 14, 16, 18, 20, 25
	Minimum number of samples required to split an internal node	2, 3, 4
	Minimum number of samples required to be at a leaf node	1, 2, 3, 4
XGBoost	Number of estimators	20, 30, 40, 50, 75, 100, 120
	Max tree depth	10, 12, 14, 16, 18, 20, 25
	Learning rate	0.01, 0.05, 0.1, 0.2
	Subsample ratio of columns when constructing trees	0.25, 0.5, 0.75, 1
	Subsample ratio of the training instances	0.25, 0.5, 0.75, 1
ANN	Number of hidden layers	1, 2, 3, 4, 5
	Number of neurons in the hidden layers (independently from each other)	from 5 to 50 with steps of 5
	Activation function in the hidden layers	sigmoid, tanh, linear, ReLU
	Dropout rate	0, 0.1, 0.2, 0.3, 0.4, 0.5

lower the AUC value is, the worse the performance of the classification algorithm in the assigned task. Since the AUC value effectively represents the performance of various machine learning algorithms, hyperparameter tuning was conducted on the basis of this metric.

The performance achieved by the developed models on *Article Set 2* was assessed using these six quality indicators. The results suggest that the effectiveness of classifying the two article groups was comparable among the different models; thus, the overall applicability of the developed model could be assessed.

3. Results

In this section, we present the results obtained during the processes of developing and evaluating the classification models. The research findings outlined in Section 3.1 were used to determine whether scientific articles could be classified on the basis of their abstracts or whether the entire text needed to be processed. In Section 3.2, the efficiencies of the different machine learning algorithms were compared. In Section 3.2, the impacts of different text preprocessing steps on the classification results were evaluated. In Section 3.3, we evaluated how well the proposed model performed when classifying a set of articles, which contained articles with a completely different topic than those of the articles in the training set used to develop the model. Finally, in Section 3.4, we provide a detailed overview of our proposed methodology and model for classifying scientific articles.

3.1. Determination of the classification input (RQ_2)

With the second research question (RQ_2), we aimed to determine whether article classification could be achieved solely by processing abstracts or whether the entire text needed to be processed. To answer this question, *Article Set 1* was classified in two ways: (i) by using only the abstracts for classification and (ii) by using the full texts of the articles as classification inputs. For a fair comparison, the hyperparameters of the machine learning algorithms were separately fine-tuned in both cases. The results of the two classification tasks are shown in Table 4.

As shown in Table 4, the tuned machine learning algorithms performed significantly worse when the classification process was based solely on the abstracts of the articles. The difference between the AUC values obtained from classifications based on abstracts and full-text classifications was examined using the Wilcoxon signed rank test. For all the analyzed machine learning algorithms, we observed a significant difference ($p \ll 0.0001$) between the abstract-based and full-text-based classifications. This difference may have arisen from the fact that abstracts, while providing concise summaries of studies, often lack the necessary methodological depth. In contrast, methodological information is embedded throughout the full text, particularly in sections such as "Methods" or "Materials and Methods", where researchers explicitly describe their approaches. Many of the distinguishing features between quantitative and qualitative research, such as statistical analyses and data collection techniques, are described in these sections, making them critical for classification. Additionally, the terminology contained in abstracts can be ambiguous, with some phrases applicable to both research paradigms, requiring additional context from the full text to ensure accurate classification results.

Since the metrics clearly indicated that a better automatic classifier could be developed when the research methods used in articles, i.e., whether quantitative or qualitative approaches were used, were classified by processing the entire text of the articles, we focused exclusively on developing machine learning models that received the full texts of the articles as their inputs.

3.2. Examining the effectiveness of the investigated machine learning algorithms (RQ_1)

Since the input dataset was determined by comparing the performance of fine-tuned machine learning models, the resulting metrics allowed us to compare the effectiveness of the tested machine learning algorithms.

The bottom half of Table 4 shows that the models built with the logistic regression and support vector machine classifiers yielded the least-favorable results. An exception to this was the sensitivity value produced by the SVM classifier, but we propose two explanations for this value. Although the results indicate that the SVM model was indeed better at categorizing qualitative articles than the other approaches were, owing to the low occurrence of positive samples in the test set, the fact that this model classified only a few articles differently than the other models did resulted in a significant improvement in this metric.

In addition, the RF and XGBoost algorithms performed slightly better than the logistic regression and SVM models did. Since the RF and XGBoost algorithms both involve tree-based learning models, we conclude that the performance of tree-based models surpasses the performance of statistical models in this particular task. This observation is in line with the results obtained by Eykens et al. (2021), who noted that a tree-based method (with gradient boosting) achieved more accurate and stable predictions related to the paper classification problem. However, they stated similar conclusions with respect to the

Table 4
Comparison between the performances of different ML algorithms in terms of whether they received the abstracts or the full text of articles as their inputs.

Article set 1 - Abstract						
Algorithm	Accuracy	Specificity	Sensitivity	F1 score	Weighted F1	AUC
LR	0.5668 (±0.1303)	0.3836 (±0.0986)	0.9088 (±0.0532)	0.5625 (±0.1188)	0.5528 (±0.1330)	0.6412 (±0.0859)
SVM	0.5940 (±0.1749)	0.4879 (±0.1810)	0.7863 (±0.0742)	0.5915 (±0.1141)	0.5952 (±0.1199)	0.6371 (±0.1017)
RF	0.5572 (±0.1342)	0.3613 (±0.1014)	0.9124 (±0.0683)	0.5513 (±0.1206)	0.5387 (±0.1064)	0.6368 (±0.1093)
XGBoost	0.5343 (±0.1438)	0.5077 (±0.1507)	0.5824 (±0.1526)	0.4428 (±0.0989)	0.4580 (±0.1076)	0.5451 (±0.1252)
ANN	0.6689 (±0.0955)	0.8855 (±0.0518)	0.4763 (±0.0861)	0.5515 (±0.1102)	0.6138 (±0.1151)	0.5809 (±0.1038)
Article Set 1 - Full text						
Algorithm	Accuracy	Specificity	Sensitivity	F1 score	Weighted F1	AUC
LR	0.8355 (±0.0487)	0.8430 (±0.0660)	0.8216 (±0.0954)	0.8228 (±0.0517)	0.8369 (±0.0480)	0.8323 (±0.0522)
SVM	0.8135 (±0.0532)	0.7574 (±0.0785)	0.9186 (±0.0664)	0.8076 (±0.0526)	0.8172 (±0.0520)	0.8380 (±0.0478)
RF	0.8581 (±0.0491)	0.8625 (±0.0675)	0.8498 (±0.0899)	0.8469 (±0.0517)	0.8592 (±0.0484)	0.8562 (±0.0511)
XGBoost	0.8523 (±0.0499)	0.8528 (±0.0684)	0.8513 (±0.0878)	0.8414 (±0.0523)	0.8537 (±0.0490)	0.8521 (±0.0512)
ANN	0.9007 (±0.0353)	0.9203 (±0.0412)	0.8638 (±0.0772)	0.8903 (±0.0396)	0.9004 (±0.0357)	0.8920 (±0.0412)

Table 5
Comparison between the performances of ML algorithms with or without using lemmatization during the preprocessing phase.

Article set 1 - Full text - No lemmatization						
Algorithm	Accuracy	Specificity	Sensitivity	F1 score	Weighted F1	AUC
LR	0.8355 (±0.0487)	0.8430 (±0.0660)	0.8216 (±0.0954)	0.8228 (±0.0517)	0.8369 (±0.0480)	0.8323 (±0.0522)
SVM	0.8135 (±0.0532)	0.7574 (±0.0785)	0.9186 (±0.0664)	0.8076 (±0.0526)	0.8172 (±0.0520)	0.8380 (±0.0478)
RF	0.8581 (±0.0491)	0.8625 (±0.0675)	0.8498 (±0.0899)	0.8469 (±0.0517)	0.8592 (±0.0484)	0.8562 (±0.0511)
XGBoost	0.8523 (±0.0499)	0.8528 (±0.0684)	0.8513 (±0.0878)	0.8414 (±0.0523)	0.8537 (±0.0490)	0.8521 (±0.0512)
ANN	0.9007 (±0.0353)	0.9203 (±0.0412)	0.8638 (±0.0772)	0.8903 (±0.0396)	0.9004 (±0.0357)	0.8920 (±0.0412)
Article Set 1 - Full text - Using lemmatization						
Algorithm	Accuracy	Specificity	Sensitivity	F1 score	Weighted F1	AUC
LR	0.8332 (±0.0511)	0.8442 (±0.0676)	0.8124 (±0.0970)	0.8198 (±0.0541)	0.8343 (±0.0504)	0.8283 (±0.0546)
SVM	0.8287 (±0.0494)	0.7750 (±0.0721)	0.9295 (±0.0640)	0.8229 (±0.0493)	0.8322 (±0.0483)	0.8522 (±0.0451)
RF	0.8693 (±0.0478)	0.8702 (±0.0621)	0.8677 (±0.0885)	0.8591 (±0.0511)	0.8704 (±0.0473)	0.8689 (±0.0509)
XGBoost	0.8715 (±0.0478)	0.8714 (±0.0629)	0.8718 (±0.0877)	0.8616 (±0.0506)	0.8726 (±0.0471)	0.8716 (±0.0505)
ANN	0.9511 (±0.0394)	0.9550 (±0.0565)	0.9438 (±0.0674)	0.9465 (±0.0422)	0.9512 (±0.0391)	0.9494 (±0.0414)

models, and importantly, that research aimed to classify papers based on their research topics and not their research methodologies. Our results highlight that tree-based methods are generally more effective in paper classification tasks not only in topic prediction cases but also when the aim is to perform categorization based on research methodologies.

Moreover, the fine-tuned neural network model consistently outperformed the models generated by other machine learning algorithms, except in terms of the sensitivity value of the SVM algorithm. The achieved quality metrics exceeding 0.9 are notably promising. Since the machine learning algorithms yielded particularly good classification results during the model development phase, we next investigated how the effectiveness of the developed models could be improved by applying some text preprocessing steps.

Examining the impact of text preprocessing (RQ_3)

Text preprocessing steps generally have significant impacts on the effectiveness of classification models. Since certain terminologies contained in various articles may differ in terms of their usage of lowercase or uppercase, converting text to lowercase can only improve the obtained classification results. Accordingly, the texts of the given articles were consistently converted to lowercase. However, an interesting research question is whether lemmatization improves or impairs the effectiveness of classification models. Lemmatization allows for the unification of various conjugated forms of a word, so we expected this preprocessing step to improve the resulting classification performance. To address this question, we developed two models on the basis of *Article Set 1*. For the first model, the texts were lemmatized before performing tokenization; for the second model, the original words were tokenized without conducting lemmatization. The effectiveness of these two ML models is compared in Table 5.

For the SVM, RF, XGBoost, and neural network models, the models developed based on lemmatized text performed better in terms of every metric than did the models developed on the basis of nonlemmatized text. The efficiencies of the logistic regression models developed based on the lemmatized and nonlemmatized texts were similar. The most significant improvement was observed for the neural network model, where the specificity and sensitivity of the model developed based on lemmatized text were 0.9550 and 0.9438, respectively, indicating very high accuracy in terms of classifying both qualitative and quantitative research-based articles. This observation is supported by the AUC value of 0.9494 yielded by the neural network, demonstrating the good performance of the applied methodology in this task. To assess whether the AUC values differed significantly among the models, we conducted a Wilcoxon signed-rank test. The test revealed significant differences between the AUC values obtained for the lemmatized and nonlemmatized datasets by all classifiers (LR: $p = 5.79e\text{-}4$, SVM: $p = 2.71e\text{-}11$, RF: $p = 3.15e\text{-}09$, XGBoost: $p = 1.30e\text{-}16$, and ANN: $p = 3.78e\text{-}88$). The Wilcoxon signed-rank test conducted on the weighted F1 scores yielded results similar to those obtained for the AUC values, confirming significant differences between the lemmatized and nonlemmatized datasets for most classifiers (LR: $p = 0.0025$, SVM: $p = 7.76e\text{-}12$, RF: $p = 7.42e\text{-}10$, XGBoost: $p = 2.03e\text{-}16$, and ANN: $p = 1.27e\text{-}92$).

On the basis of the results, we can conclude that by applying appropriate text preprocessing steps, a highly accurate classification model could be developed for the defined task. However, how corpus-specific the developed models are, i.e., whether they can be generally applied to other text corpora, remains an important question. In the next subsection, we examine this question.

3.3. The application of the proposed model to another corpus (RQ_3)

The ML models fine-tuned based on *Article Set 1* were saved and then applied to classify the articles contained in *Article Set 2*. The classification results obtained by the models trained on the lemmatized and nonlemmatized versions of *Article Set 2* are presented in Table 6.

Table 6
Evaluation of the performance achieved by ML algorithms based on different corpora.

Article Set 2 - Full text - No lemmatization						
Algorithm	Accuracy	Specificity	Sensitivity	F1 score	Weighted F1	AUC
LR	0.9055 (±0.0379)	0.9082 (±0.0388)	0.7712 (±0.0567)	0.6059 (±0.0427)	0.9356 (±0.0228)	0.8397 (±0.0318)
SVM	0.8904 (±0.0153)	0.8916 (±0.0156)	0.8293 (±0.0308)	0.5868 (±0.0169)	0.9269 (±0.0090)	0.8604 (±0.0172)
RF	0.9284 (±0.0133)	0.9309 (±0.0140)	0.8075 (±0.0833)	0.6373 (±0.0230)	0.9494 (±0.0078)	0.8692 (±0.0397)
XGBoost	0.9357 (±0.0105)	0.9384 (±0.0110)	0.8025 (±0.0574)	0.6499 (±0.0209)	0.9537 (±0.0062)	0.8704 (±0.0276)
ANN	0.8988 (±0.0189)	0.9001 (±0.0198)	0.8336 (±0.0674)	0.5996 (±0.0209)	0.9318 (±0.0111)	0.8669 (±0.0304)
Article Set 2 - Full text - Using lemmatization						
Algorithm	Accuracy	Specificity	Sensitivity	F1 score	Weighted F1	AUC
LR	0.9184 (±0.0264)	0.9220 (±0.0270)	0.7428 (±0.0431)	0.6166 (±0.0342)	0.9432 (±0.0156)	0.8324 (±0.0233)
SVM	0.8705 (±0.0200)	0.8706 (±0.0207)	0.8656 (±0.0550)	0.5702 (±0.0165)	0.9151 (±0.0120)	0.8681 (±0.0240)
RF	0.9205 (±0.0478)	0.9223 (±0.0621)	0.8294 (±0.0885)	0.6267 (±0.0510)	0.9447 (±0.0473)	0.8758 (±0.0509)
XGBoost	0.9221 (±0.0134)	0.9234 (±0.0138)	0.8601 (±0.0624)	0.6330 (±0.0215)	0.9458 (±0.0079)	0.8917 (±0.0307)
ANN	0.9205 (±0.0138)	0.9233 (±0.0141)	0.7782 (±0.0178)	0.6230 (±0.0233)	0.9446 (±0.0082)	0.8508 (±0.0113)

As shown in Table 6, the accuracy, specificity, and AUC values of the LR, SVM, RF, and XGBoost models improved when classification was performed on the basis of *Article Set 2*. This trend could be observed in the classification results produced by both articles preprocessed with lemmatization and those without lemmatization preprocessing. Thus, the developed models exhibited good generalizability and did not overfit *Article Set 1*. However, we observed significant decreases in the sensitivity values of all the models. This decrease can be attributed to the fact that out of the 557 articles contained in *Article Set 2*, only 11 belong to the qualitative research category. If a model misclassified a qualitative article, this metric significantly decreased. This effect was also reflected in the F1 score values.

In contrast with the other models, for the neural network model, we observed that the performance metrics were consistently worse when classifying *Article Set 2* than when classifying *Article Set 1*. This trend indicates that the generalizability of the ANN model was less satisfactory than that of the other models and that the model overfitted the classification task based on *Article Set 1*. To reduce the degree of overfitting suffered by the ANN model, the performance of several less complex models was also tested. Reducing the complexity of the neural network model did not improve its generalizability, and the less complex models achieved similar results when classifying articles in *Article Set 2*, as did the fine-tuned original ANN model. Notably, augmenting the training dataset could have improved the generalizability of the model.

3.4. The proposed classification methodology and model for classifying scientific articles on the basis of their research methods

Upon reviewing all the research results, we can conclude that applying lemmatization as a text preprocessing operation improved the achieved article classification accuracy. The model with the best quality indicators and generalizability was the XGBoost model, which contained 30 decision trees, each with a maximum depth of 14. The XGBoost model achieved an accuracy of 0.8715 and an AUC of 0.8716 on the lemmatized *Article Set 1*. Applying the same model to *Article Set 2* resulted in an accuracy of 0.9221 and an AUC of 0.8917. The Wilcoxon rank-sum test revealed a statistically significant difference ($p \ll 0.0001$) when the AUC values of the XGBoost model were compared with those of the logistic regression, support vector, and neural network models. However, no significant difference was found between the AUC values of the random forest and XGBoost models, with $p = 0.515$ and $p = 0.516$ for the classification results obtained on *Article Set 1* and *Article Set 2*, respectively. However, when comparing the evaluation results of the random forest and XGBoost models, we observed that the XGBoost model exhibited slightly better performance and greater stability. On this basis, we can determine that the developed XGBoost classifier consistently achieves outstanding article classification results and that the generalizability of the model is adequate. Additionally, XGBoost does not require significant computational resources, as it is optimized for efficiency through parallel processing and built-in regularization techniques, making it well suited for large-scale datasets even when hardware resources are limited.

3.5. Comparison between the proposed classification model and transformer-based large language models

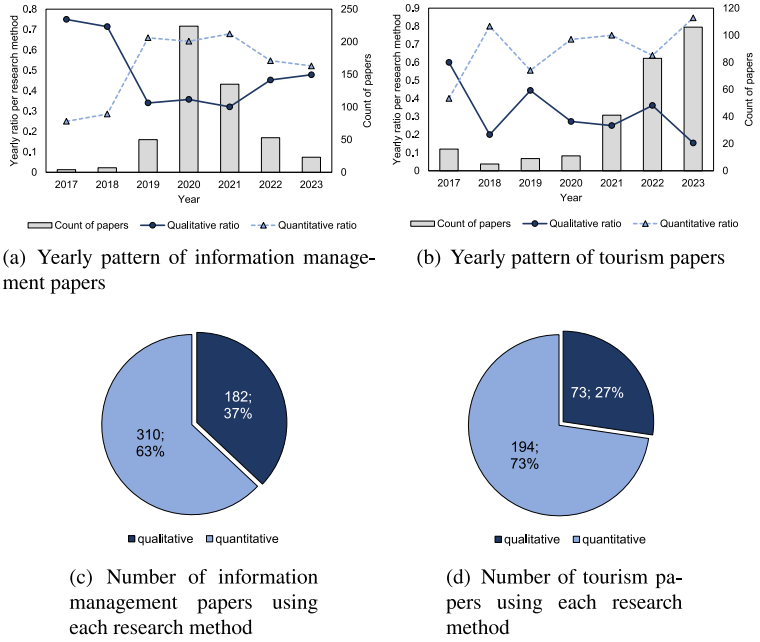
The proposed XGBoost model demonstrated good performance. However, it is worth comparing its performance with that of multiple Large Language Models (LLMs) as well. To address this question, we compared the performance of our model with that of artifical intelligence (AI) models, i.e., ScholarGPT (OpenAI, 2025) and NotebookLM (Google Research, 2025). In this comparison, the AI models were tasked with determining whether the uploaded articles employed quantitative or qualitative research methods. The analysis was conducted on both *Article Set 1* and *Article Set 2*. Owing to memory limitations, both models were unable to process the entire analysis at once. Consequently, we had to divide the article sets into chunks containing 10 articles each and then aggregate the partial results. The confusion matrices produced by the tested transformer models can be found in Table 7.

As we can see, there were articles for which the AI models could not clearly determine the research methodologies applied in the corresponding study, and an 'Uncertain' category was introduced for these cases. Comparing ScholarGPT and NotebookLM, we observe that this issue occurred in a significant percentage of the articles, leading to a substantial decline in model performance compared with that of our developed classification model. The accuracies of ScholarGPT were 0.4018 on *Article Set 1* and 0.5135 on *Article Set 2*, whereas its sensitivity values were 0.3625 and 0.8182, and its specificity scores were 0.4228 and 0.5073 for *Article Set 1* and *Article Set 2*, respectively. In the case of NotebookLM, the corresponding values were accuracies of 0.6987 and 0.8456, sensitivities of 0.4500 and 0.1818, and specificities of 0.8322 and 0.8590 for *Article Set 1* and *Article Set 2*, respectively. These results clearly demonstrate that our fine-tuned XGBoost-based classification model significantly outperformed the investigated AI models in terms of both predictive performance and computational efficiency.

A possible reason is that AI models primarily consider citations when categorizing research. First, if an AI model recognizes that a study is built on quantitative methods, yet the authors cite a significant number of qualitative articles, it may be difficult to determine the proper classification of the research (qualitative or quantitative). Second, if a study is qualitative but cites an excessive number of sources where the main text includes standalone years, year-page references, or numerical citation formats, the AI may misinterpret it as quantitative research. In our view, AI models recognize that the presence of numbers should be considered when determining whether a study is quantitative. However, they are unable to distinguish between publication years, page

Table 7
The confusion matrices of the classification results provided by the ScholarGPT and NotebookML models.

ScholarGPT	Article set 1			Article set 2		
	Predicted quantitative	Predicted qualitative	Uncertain	Predicted quantitative	Predicted qualitative	Uncertain
True Quantitative	63	8	78	277	241	28
True Qualitative	9	29	42	2	9	0
NotebookLM	Article Set 1			Article Set 2		
	Predicted Quantitative	Predicted Qualitative	Uncertain	Predicted Quantitative	Predicted Qualitative	Uncertain
True Quantitative	124	3	22	469	1	76
True Qualitative	7	36	37	3	2	6



(a) Yearly pattern of information management papers

(b) Yearly pattern of tourism papers

(c) Number of information management papers using each research method

(d) Number of tourism papers using each research method

Fig. 4. Overall and yearly ratios by research method. The proposed pretrained classifier was used for Article Set 3. Subfigures a and b show the yearly number of papers published (bars) and the ratio of the research methods used (solid and dashed lines) for each topic. Pie charts represent the distribution of the research methods used for each research topic (subfigures c and d).

numbers, and actual research results. As noted by Williams (2024), we also emphasize that AI models' interpretation of data in qualitative and quantitative research should be approached with caution.

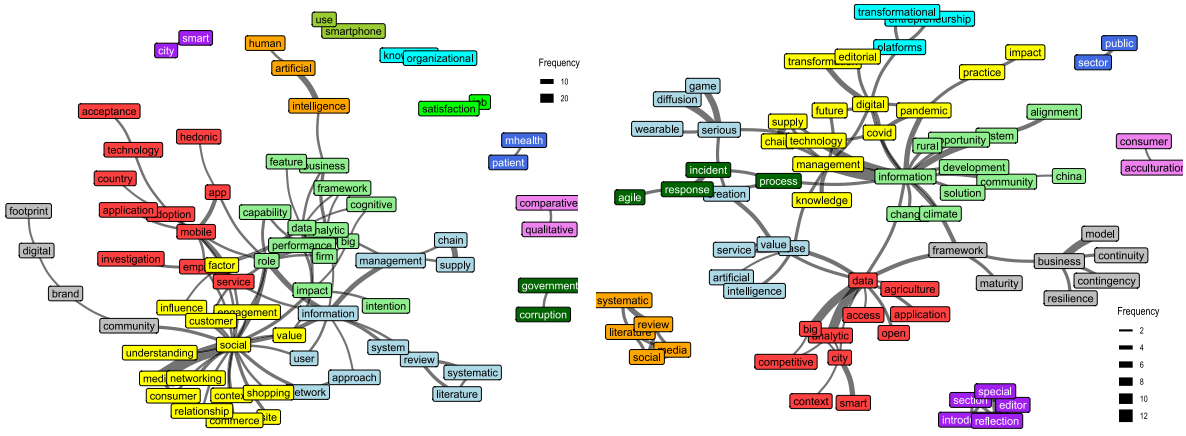
4. Application example

The proposed classifier was also tested on *Article Set 3* to investigate its generalizability and applicability to different research fields. All the papers were classified by the model, and the two research fields included in *Article Set 3* were analyzed while considering their characteristics in terms of the research methods used. Fig. 4(a) and Fig. 4(b) show the yearly patterns of the ratios of the research methods used in different types of papers. Fig. 4(c) and Fig. 4(c) show the overall ratios for each research method.

In Figs. 4(a-b), the bars represent the frequencies of the papers published by year, and the solid and dashed lines denote the ratios of the numbers of qualitative and quantitative studies, respectively. The pie charts (Figs. 4(c-d)) show the overall percentages of studies following different research methods. After applying the developed classifier, it was clear that the two research fields had different longitudinal pattern ratios between quantitative and qualitative methods. In the information management field, the quantitative and qualitative approaches were applied in approximately equal numbers of papers (approximately 50%–50%);. However, in the tourism management field, the applications of the two research methods considerably differed over time (Figs. 4(a–b)). On the other hand, the articles published in the two research fields did not significantly differ in terms of the overall research methods used (Figs. 4(c-d)).

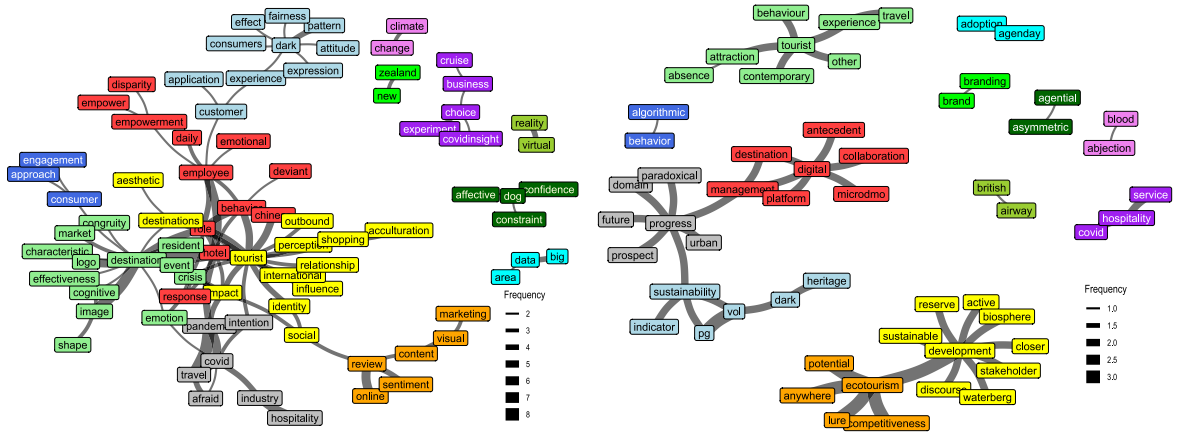
Additionally, the longitudinal patterns, the overall main topics in each research field and the research method used to investigate each main topic could all be analyzed with the developed classifier. Figs. 5(a-b) show the word co-occurrence networks produced for information management papers in which quantitative (Fig. 5(a)) and qualitative (Fig. 5(b)) approaches were used. Similarly, Fig. 5(c-d) illustrates the word co-occurrence networks produced for tourism management papers in which quantitative and qualitative approaches were used.

To construct the co-occurrence networks, the titles of the papers were collected and preprocessed. The text preprocessing steps included lowercase transformation, lemmatization, special character processing and stop-word removal. Afterward, the co-occurrence values of the



(a) Information management papers predicted as using quantitative methods

(b) Information management papers predicted as using qualitative methods



(c) Tourism papers predicted as using quantitative methods

(d) Tourism papers predicted as using qualitative methods

Fig. 5. Co-occurrence networks established from paper titles. The nodes represent the keywords, and the edges denote the frequencies of the keyword pairs occurring in the same title. The colors represent the results of performing community detection via the Leiden algorithm. The titles were preprocessed using lowercase transformation, lemmatization, special character removal, and stop-word removal.

words were calculated, where co-occurrence represents the frequency with which the two words occurred in the same title. In other words, the nodes of the networks represent the unique words occurring in the corpus, and the weights of the edges indicate how many times the two connected words occurred within the same paper title. After constructing the network, a community detection algorithm was applied to cluster the words and detect the densely interconnected subgraphs, i.e., topics. Figs. 5(a-b) show that in the case of the information management papers, the quantitative papers mostly addressed social networks, mobile applications, information systems, big data and digital footprints, whereas the qualitative studies mainly focused on smart cities, business models, knowledge management, and AI. In tourism research (Figs. 5(c-d)), the main topics of the quantitative papers were tourist perception, pandemic effects, destination management and tourism marketing, whereas the qualitative studies addressed ecotourism, sustainable development, digital management and the tourist experience.

5. Discussion

In this work, the results showed that an automated machine learning model could be built and used to predict and classify the research methods used in research papers (RQ_1). In the presented practical example, the specific approaches proved to be more effective and accurate than the generic approaches. The classification performance improved only slightly when text lemmatization was used. Therefore, the results suggest that while lemmatization can improve the results, selecting an adequate model is more important. Surprisingly, with the lemmatized corpus, the generic algorithms achieved better performance than the specific algorithms did because of the cleaner corpus and simplified structure. This also suggests that both specific and generic methods can be applied to solve the defined problem. Nevertheless, manually labeling the papers contained in the training and test data can bias the resulting model performance; in this case, the proposed models still performed well. These results are very usable by scholars since, as noted, more effective and accurate scientometric analyses can be performed by emphasizing the model selection process. The developed automatic classifier also accelerates the scientometric analysis procedure, which increases the motivation to conduct research paper classification-based studies (CO_1) and the speed at which new research results can be disseminated (CO_2). The use of the classifier also extends scientometric analyses to include research method classification (CO_3) and eliminates subjective decision-making steps, which increases the validity and reproducibility of the research (classification) results (CO_4). This is very important because a research method-based classification approach for scientific papers was not previously available.

With respect to the adequacy of the input data for the proposed classifier, the results showed that significantly better results could be

achieved if not only the abstracts but also the entire bodies of the papers were used as the corpus for training (RQ_2). This suggests that it is not worth considering only abstracts as a corpus when research paper classifiers are developed; rather, the entire bodies of the papers should be used to achieve better results. These results also indicate that the state-of-the-art solutions might not use an adequate strategy during the paper classification stage since they mostly rely on abstracts and do not consider the full texts of the examined papers. The abstracts might not precisely reflect the research methods used in the papers, so researchers are encouraged to investigate the details of the papers to determine which methods were used for their analyses.

Finally, the proposed model and vocabulary were applied to another set containing tourism management articles, and the results showed that the developed classifier performed well on data derived from different research fields (RQ_3). In addition to the overall good performance, the differences in the sensitivity values indicate that the proposed model can still be improved to more efficiently handle imbalanced datasets. Although the overall AUC was satisfactory, achieving increased robustness in cases with imbalanced classes can be a direction for future research. The proposed classifier could successfully identify the main research topics within the two article sets, i.e., information management and tourism research, and the topicwise differences between qualitative- and quantitative-based papers were revealed. In the case with information management papers, the quantitative papers focused mostly on social networks, mobile applications, information systems, big data and digital footprints, whereas the qualitative studies focused mainly on smart cities, business models, knowledge management, and artificial intelligence. After applying the same classifier to the tourism dataset, we found that the quantitative papers mainly addressed tourist perceptions, pandemic effects, destination management and tourism marketing, whereas the qualitative studies focused mainly on ecotourism, sustainable development, digital management and the tourist experience. When the ratios of quantitative-qualitative studies over time in both research fields were investigated, we found considerable differences between the two fields. In information management, the numbers of quantitative and qualitative papers tended to be approximately equal (approximately 50%); however, in tourism management, a considerably different trend was observed (in 2023, ~80% of the papers were quantitative, whereas ~20% of the papers were qualitative).

On the one hand, this application and analysis prove that the proposed model and the predefined vocabulary are generalizable and can be used in different research fields without modifications or adjustments. This result also highlights the notion that research papers have common parts, such as their research methodologies, which are independent of the research field, emphasizing the opportunity to develop generalizable automatic classifiers. Compared with the existing approaches, the proposed method is a significant contribution to the field, since the previous works did not analyze how their models performed depending on the context (research field) of the test corpus. Nevertheless, it is worth mentioning that in future work, a comparison between the document-term matrix approach and embedding techniques would be valuable since this step could further decrease the level of human involvement in the modeling process. On the other hand, the application example demonstrates that the proposed automatic classifier can be used to quickly obtain valuable information and important differences in multiple research fields.

6. Summary and conclusion

In this paper, we propose an accurate and automatic classification model for analyzing and predicting the research methodologies used in scientific articles. The contributions and implications (Is) of this paper are presented in Table 8.

In this paper, the results showed that the research methodology classification problem can be automated with machine learning algorithms, yielding satisfactory accuracy. The specific methods outperformed the generic methods in terms of classification accuracy, achieving accuracies exceeding 90%. For the scientific community, the proposed model can be used to automate scientometric analyses (C_1). This is extremely important because it provides the opportunity to analyze and validate new research questions that cannot be analyzed without the proposed model. For example, we can understand how the utilized research methodologies evolve over time or what differences can be observed among the research methodologies used by researchers in different fields. In addition, the results imply that through automated methods, the time needed to disseminate research results can be significantly decreased (I_1), which is valuable for researchers and policy makers. Although the overall performance of the proposed model was satisfactory, the process of manually labeling the training and testing sets and making human decisions for vocabulary building introduced biases to the analysis. It would be beneficial in future work to extend the analysis with word embedding methods to decrease the level of human involvement in the process.

In addition, the results showed that the model quality was significantly higher when the entire text of the papers was used than when only the abstracts were used. Thus, using only abstracts in scientometric analyses is not suggested (C_2). As an important implication for editors and scholars, the results indicate that abstracts do not comprehensively describe the key elements of papers, and considering only abstracts in scientometric analyses may lead to false conclusions (I_2). Moreover, the results suggest that editors should review abstracts more carefully.

Finally, the proposed classification model is generalizable; however, generic models exhibit more consistent performance than do neural networks. We showed that the proposed model and vocabulary are applicable to different research fields (C_3). This contribution also indicates that research papers have common parts, such as their employed methodologies, that do not depend on the field. The parts that do not depend on the field can be identified, and the corresponding classification task can be automated using the proposed model. Nevertheless, the results indicated that the proposed model might be sensitive to imbalanced datasets; therefore, a possible future direction could be improving the robustness of the model with respect to imbalanced data. For the practical support of scholars, the code and vocabulary of the developed model are made available as support material (I_3). As all the contributions and implications show, the results of this work are beneficial for the scientific community since the proposed method provides an opportunity to conduct not only topic-focused work but also research method classification tasks that were not previously available. Users can employ the best-performing model on the basis of our results, alongside the provided glossary on any corpus derived from many research fields. These benefits will improve the efficiency and quality of future scientometric analyses.

7. Limitations and future work

While we have made significant progress in terms of automating the research methodology classification process, some limitations should be considered. One key limitation is the reliance on machine learning algorithms, which may introduce biases or inaccuracies on the basis of the provided training data. As a potential bias, the manual labeling process related to the research methods used in the papers contained within the training and test data can lead to model distortions when fitting model predictions. Additionally, the effectiveness of the classification model may vary depending on the quality and quantity of the input data, highlighting the need for robust data collection processes. Another limitation is the generalizability of the model across different research fields, as specific adaptations or additional features may be required for models in certain fields to ensure the acquisition of accurate classification results. However, the proposed model attained satisfactory performance on different datasets, and the sensitivity value differences suggest that the model was still sensitive to imbalanced

Table 8
Summary of the results, contributions and implications.

Research question	Research result	Contribution	Implication
RQ_1: Can the process of classifying the research methods employed in articles be automated using machine learning methods? If so, which machine learning algorithm is most efficient for classifying the research methodologies used in scientific articles, and what classification accuracy can be achieved using different machine learning algorithms?	The research methodology classification problem can be automated by machine learning algorithms. The fine-tuned XGBoost model proved to be satisfactory in terms of classifying research papers based on their research methodologies. This model achieved better performance than other machine learning algorithms did (the quality metrics were all greater than 0.9).	C_1: Scientometric analyses can be supported by the automated research methodology classification method.	I_1: The time required to disseminate research based on scientometric analyses can be significantly decreased.
RQ_2: Can article classification be performed solely based on abstracts, or is it necessary to process the entire text of the examined articles for classification purposes? Furthermore, what text preprocessing steps should be performed before conducting classification?	Significantly higher model quality metrics were achieved when the entire text of the papers was used for classification. Converting the text of each article to lowercase and then lemmatizing it improved the accuracy of the model.	C_2: We showed that using only the abstracts for classification was not sufficient, and the full text should be used for classification.	I_2: Due to length constraints, abstracts generally do not provide sufficiently detailed information regarding the research objectives, methodology, results, and conclusions, thus failing to give the reader a comprehensive overview of the content of an article. Considering only the abstract in such an analysis is superficial, and important information may be missed with this approach.
RQ_3: Should corpus-dependent models be created, and how generalizable is the proposed classification model?	The classification model is generalizable. The generic models exhibited more consistent performance than the specific methods did.	C_3: With the proposed model and vocabulary, papers related to different research fields could be classified and analyzed.	I_3: The code and vocabulary are available for interested researchers.

datasets. Finally, the proposed model was trained and tested on various datasets, including hundreds of papers, but the efficiency and computation time of the model were not analyzed for large databases. Therefore, scaling of the model to efficiently handle large volumes of research papers remains a challenge that needs to be addressed.

There are several directions for future research that can enhance and expand upon our findings. One potential direction is to explore the integration of natural language processing techniques for extracting more nuanced information from research papers, allowing for a deeper understanding of the methodologies employed. Additionally, investigating the impacts of different feature selection methods and model architectures on the resulting classification performance could provide valuable insights for optimizing the classification process. In addition to improvements with more efficient feature selection approaches, the integration of transformer-based models and the application of domain adaptation and modern embedding techniques in prediction tasks are valuable future extensions of the proposed method. This could also support further increasing the generalizability of the findings. The applications of the model and explorations of new research methods are also valuable opportunities for further analysis. Furthermore, conducting comparative studies with other automated classification models and refining the proposed model on the basis of feedback and real-world applications can help validate its effectiveness and reliability. Collaborating with domain experts to tailor the model to specific research fields and continuously updating the model with new data to ensure its relevance and accuracy are also crucial areas for future exploration.

CRediT authorship contribution statement

Zsolt T. Kosztyán: Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Data curation, Conceptualization. **Tünde Király:** Writing – original draft, Resources, Project administration, Data curation. **Tibor Csizmadia:** Writing – review & editing, Writing – original draft, Validation, Investigation, Data curation, Conceptualization. **Attila Imre Katona:** Writing – review & editing, Writing – original draft, Validation, Methodology, Data curation, Conceptualization. **Ágnes Vathy-Fogarassy:** Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Code availability

The code is available at https://github.com/kzst/article_classifier.

Funding

This research is supported by the Research Centre at the Faculty of Business and Economics (PE-GTK-GSKK A095000000-7) of the University of Pannonia (Veszprém, Hungary). Project no. K 143482 has been implemented with support provided by the Ministry of Culture and Innovation of Hungary from the National Research, Development and Innovation Fund, financed under the K_22 “OTKA” funding scheme.

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

See Table A.9.

Table A.9
List of terms determined on the basis of expert knowledge.

Accuracy	Distribution	Meta-Analysis	Replicability
Action Research	Double-Barreled Question	Micro-Ethnography	Representativeness
Administrative Data	Double-Barreled Question	Missing Completely at Random	Respondent
Affective Measure	Double-Blind	Missing Data	Response Categories
Aggregate	Duration Model	Misspecification	Retrospective Consent
Alpha Level	Effect Size	Mixed Method	Retrospective Look
ANCOVA	Endogeneity	Mode	Rigorous Research
Anonymity	Estimation	Monotonic Trend	Robustness
ANOVA	Ethnographic Decision Model	Monte Carlo	Sample
Association	Ethnography	Multidimensional	Sampling
Attrition	Evaluation Apprehension	Multifactor	Scaled Score
Average	Event History Analysis	Multicollinearity	Scatter Plot
Axiom	Exogeneity	Multidimensional	Scattergram
Behavioral Categories	Experimental Control	Multinomial	Scatterplot
Bell-Shaped Curve	Experimental Design	Multivariate	Scenario Analysis
Benchmarking	Expert Panel	Mutually Exclusive	Scenario Planning
Beta Level	Expert study	Narrative Research	Secondary Data
Between-Subjects Design	Explanatory Analysis	Natural Experiment	Selective Observation
Bias	Exploratory Data Analysis	Naturalistic Observation	Semantic Differential Scale
Bootstrapping	Exploratory Study	Negative Association	Sensitivity Analysis
Case Study	Extrapolation	Nominal	Sign Test
Categorical Data	Factor Analysis	Nonlinear	Significance
Causal Analysis	Field experiment	Nonparametric	Simulation
Causal Inference	Field Research	Nonresponse Error	Single Blind
Causal Relationship	Focus Group	Nonsampling Error	Skewness
Causality	Focus Study	Nonsignificant	Slope
Ceiling	Forecasting	Norm	Social Desirability
Census	Friedman Test	Observed Value	Social Network
Central Limit Theorem	Fuzzy	Open Question	Sociogram
Central Tendency	Game Theory	Order Effect	Spurious Relationship
Chart	Generalized Linear Mixed Model	Ordinal	Standard Error
Chi Square	Generalized Linear Model	Outlier	Standardization
Chi Squared	Grounded Theory	Oversampling	Standardized
Closed questions	Hawthorne Effect	Overt Observation	Standardization
Cluster Analysis	Heterogeneity	P Value	Standardized
Coefficient	Heteroskedastic	Paired Comparison	Statistic
Cognitive Interviewing	Hierarchical Linear Modeling	Panel Study	Statistical
Cohort	Hierarchical Model	Participant Observation	Stratification
Comparability	Histogram	PASTA Analysis	Structural Equation Modeling
Composite Reliability	Hypothesis	Path Analysis	Structural Model
Confidence	Imputed Response	PCA	Structured Observation
Confidentiality	In-depth Interviewing	Percentile	Subsample
Consistency	Independent Group Design	Phenomenology	Subsample
Constant	Indicator	Pilot Study	Survey Research
Contamination	Indirect Effect	Point Estimate	T Test
Content Analysis	Inductive Method	Positively Associated	Test-Retest Reliability
Context Conditionality	Informed consent	Power	Thematic Analysis
Context Sensitivity	Insiderness	Pretest	Time-Invariant
Control Group	Instrument Error	Precision	Time Series
Controlled Experiment	Intent to Treat	Presumptive Consent	Time Variant
Controlled Observation	Inter-Observer Reliability	Pretest	Treatment Effect
Correlation	Interrater reliability	Primary data	Treatment offered
Correlational	Interaction Effect	Principal Component Analysis	Treatment on the Treated
Counterbalancing	Intercept	Probability	Triangulation
Covariable	Interval	Probit Model	Two-Tailed
Covariance	Interview	Psychometric Property	Type 1 Error
Covariate	Jackknife Technique	Q Methodology	Type 2 Error
Coverage	Kruskal Wallis Test	Qualitative	Type I Error
Covert Observation	Kurtosis	Quantile	Type II Error
Critical Incident Technique	Laboratory Experiment	Quantitative	Typology
Cronbach's Alpha	Latent Growth Model	Quartile	Unbiased
Cross-Sectional Data	Least Squares	Quasiexperiment	Unconditional Longitudinal Model
Cross Tabulation	Likert Scale	Quasiexperimental	Unconditional Longitudinal Study
Crossover Design	Link Analysis	Questionnaire	Undersampling
Curvilinear	Log-Linear Analysis	R Squared	Unfolding Question
Dashboard	Logit Model	Random Effects Model	Unit of Analysis
Data Analysis	Loglinear Analysis	Random Error	Univariate
Data Mining	Main Effect	Random Selection	Unstructured Observation
Data Reduction	Margin of Error	Randomization	Validity
Deduction	Markov Chain	Randomization	Variable
Deductive Method	Matched pairs	Range	Variance
Degrees of Freedom	Maximum Likelihood	Rank Order	Weighted Score
Delphi	MDS	Rate	Weighting
Dendrogram	Mean	Rating Scale	xij
Deviation	Measurement Error	Ratio	yij
Discriminant Analysis	Median	Regression	Z Score
Dispersion	Meta-Analysis	Repeated Measures	Z Test

Appendix B. List of symbols and abbreviations

Glossary	
AI	artificial intelligence
ANN	artificial neural network
AUC	area under the receiver operating characteristic curve
BERT	Bidirectional Encoder Representations from Transformers
C	contribution
CA	cause
CO	consequence
COVID	coronavirus disease
CNN	convolutional neural network
DL	Deep Learning
FN	false negative
FP	false positive
GPT	Generative Pre-trained Transformers
I	implication
LDA	latent Dirichlet allocation
LLM	Large Language Model
LR	logistic regression
ML	machine learning
SVM	support vector machine
PR	problem
RF	random forest
RQ	research question
RNN	recurrent neural network
ROC	receiver operating characteristic curve
TN	true negative
TP	true positive
XGBoost	Extreme Gradient Boosting

Data availability

Data will be made available on request.

References

Al Fahoum, A., 2023. Enhanced cardiac arrhythmia detection utilizing deep learning architectures and multi-scale ECG analysis. Tuijin Jishu/ J. Propuls. Technol. 44 (6), 5539–5554.

Atkinson, C.F., 2024. Cheap, quick, and rigorous: Artificial intelligence and the systematic literature review. Soc. Sci. Comput. Rev. 42 (2), 376–393. http://dx.doi.org/10.1177/08944393231196281.

Chowdhury, S., Schoen, M.P., 2020. Research paper classification using supervised machine learning techniques. In: 2020 Intermountain Engineering, Technology and Computing. IETC, IEEE, pp. 1–6. http://dx.doi.org/10.1109/IETC47856.2020.9249211.

Churruca, K., Ellis, L.A., Pomare, C., Hogden, A., Bierbaum, M., Long, J.C., Olekalns, A., Braithwaite, J., 2021. Dimensions of safety culture: a systematic review of quantitative, qualitative and mixed methods for assessing safety culture in hospitals. BMJ Open 11 (7), e043982. http://dx.doi.org/10.1136/bmjopen-2020-043982.

Colorado State University, 2024. Colorado State University Glossary. URL https://writing.colostate.edu/guides/pdfs/guide90.pdf. (Last accessed: 09 March 2025).

Daradkeh, M., Abualigah, L., Atalla, S., Mansoor, W., 2022. Scientometric analysis and classification of research using convolutional neural networks: A case study in data science and analytics. Electronics 11 (13), 2066. http://dx.doi.org/10.3390/electronics11132066.

de la Torre-López, J., Ramírez, A., Romero, J.R., 2023. Artificial intelligence to automate the systematic review of scientific literature. Computing 105 (10), 2171–2194. http://dx.doi.org/10.1007/s00607-023-01181-x.

De Santis, E., Martino, A., Ronci, F., Rizzi, A., 2025. From bag-of-words to transformers: A comparative study for text classification in healthcare discussions in social media. IEEE Trans. Emerg. Top. Comput. Intell. 9 (1), 1063–1077. http://dx.doi.org/10.1109/TETCI.2024.3423444.

Eykens, J., Guns, R., Engels, T.C., 2021. Fine-grained classification of social science journal articles using textual data: A comparison of supervised machine learning approaches. Quant. Sci. Stud. 2 (1), 89–110. http://dx.doi.org/10.1162/qss_a_00106.

Golan, R., Reddy, R., Muthigi, A., Ramasamy, R., 2023. Artificial intelligence in academic writing: a paradigm-shifting technological advance. Nat. Rev. Urol. 20 (6), 327–328. http://dx.doi.org/10.1038/s41585-023-00746-x.

Google Research, 2025. NotebookLM: AI-powered Research Assistant. https://notebooklm.google.com/. (Last accessed: 14 March 2025).

Gündoğan, E., Kaya, M., 2020. Research paper classification based on Word2vec and community discovery. In: 2020 International Conference on Decision Aid Sciences and Application. DASA, IEEE, pp. 1032–1036. http://dx.doi.org/10.1109/DASA51403.2020.9317101.

Hanyurwimfura, D., Bo, L., Njagi, D., Dukuzumuremyi, J.P., 2014. A centroid and relationship based clustering for organizing. Int. J. Multimed. Ubiquitous Eng. 9 (3), 219–234. http://dx.doi.org/10.14257/ijmue.2014.9.3.21.

Harzing, A.-W., Adler, N.J., 2016. Disseminating knowledge: From potential to reality-new open-access journals collide with convention. Acad. Manag. Learn. Educ. 15 (1), 140–156. http://dx.doi.org/10.5465/amle.2013.0373.

Hoppe, F., Dessì, D., Sack, H., 2021. Deep learning meets knowledge graphs for scholarly data classification. In: Companion Proceedings of the Web Conference 2021. pp. 417–421. http://dx.doi.org/10.1145/3442442.3451361.

Hutson, M., 2022. Could AI Help You to Write Your Next Paper?. Nature Publishing Group, http://dx.doi.org/10.1038/d41586-022-03479-w.

Kandimalla, B., Rohatgi, S., Wu, J., Giles, C.L., 2021. Large scale subject category classification of scholarly papers with Deep Attentive Neural Networks. Front. Res. Metrics Anal. 5, 600382. http://dx.doi.org/10.3389/frma.2020.600382.

Katona, A., Vathy-Fogarassy, Á., Kosztyán, Z.T., Csizmadia, T., Király, T., 2025. Automated Research Methodology Classification Using Machine Learning. figshare, http://dx.doi.org/10.6084/m9.figshare.c.7735013.v1, URL https://figshare.com/collections/Automated_Research_Methodology_Classification_Using_Machine_Learning/7735013/1.

Khadhraoui, M., Bellaaj, H., Ammar, M.B., Hamam, H., Jmaiel, M., 2022. Survey of BERT-base models for scientific text classification: COVID-19 case study. Appl. Sci. 12 (6), 2891. http://dx.doi.org/10.3390/app12062891.

Kim, S.-W., Gil, J.-M., 2019. Research paper classification systems based on TF-IDF and LDA schemes. Human- Centric Comput. Inf. Sci. 9, 1–21. http://dx.doi.org/10.1186/s13673-019-0192-7.

Lukito, J., Chen, B., Masullo, G.M., Stroud, N.J., 2024. Comparing a BERT classifier and a GPT classifier for detecting connective language across multiple social media. In: Al-Onaizan, Y., Bansal, M., Chen, Y.-N. (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Miami, Florida, USA, pp. 19140–19153. http://dx.doi.org/10.18653/v1/2024.emnlp-main.1067, URL https://aclanthology.org/2024.emnlp-main.1067/.

Mendoza, O., Kusa, W., El-Ebshihy, A.M., Wu, R., Pride, D., Knoth, P., Herrmannova, D., Piroi, F., Pasi, G., Hanbury, A., 2022. Benchmark for research theme classification of scholarly documents. In: # PLACEHOLDER_PARENT_METADATA_VALUE#, vol. 29, Association for Computational Linguistics, pp. 253–262. http://dx.doi.org/10.34726/4521.

OpenAI, 2025. ScholarGPT: AI-powered research assistant. https://platform.openai.com/. (Last accessed: 14 March 2025).

Raja, H., Munawar, A., Mylonas, N., Delsoz, M., Madadi, Y., Elahi, M., Hassan, A., Serhan, H.A., Inam, O., Hernandez, L., et al., 2024. Automated category and trend analysis of scientific articles on ophthalmology using large language models: development and usability study. JMIR Form. Res. 8 (1), e52462. http://dx.doi.org/10.2196/52462.

Rivest, M., Vignola-Gagné, E., Archambault, É., 2021. Level classification of scientific publications: A comparison of deep learning, direct citation and bibliographic coupling. PloS One 16 (5), e0251493. http://dx.doi.org/10.1371/journal.pone.0251493.

Salatino, A., Osborne, F., Motta, E., 2022. Cso classifier 3.0: a scalable unsupervised method for classifying documents in terms of research topics. Int. J. Digit. Libr. 1–20. http://dx.doi.org/10.1007/s00799-021-00305-y.

Shu, F., Ma, Y., Qiu, J., Larivière, V., 2020. Classifications of science and their effects on bibliometric evaluations. Scientometrics 125, 2727–2744. http://dx.doi.org/10.1007/s11192-020-03701-4.

Shushkevich, E., Alexandrov, M., Cardiff, J., 2023. Improving multiclass classification of fake news using bert-based models and CHATGPT-augmented data. Inventions 8 (5), 112. http://dx.doi.org/10.3390/inventions8050112.

Sulyok, J., Fehérvölgyi, B., Csizmadia, T., Katona, A.I., Kosztyán, Z.T., 2023. Does geography matter? Implications for future tourism research in light of COVID-19. Scientometrics 128 (3), 1601–1637. http://dx.doi.org/10.1007/s11192-022-04615-z.

Taheriyan, M., 2011. Subject classification of research papers based on interrelationships analysis. In: Proceedings of the 2011 Workshop on Knowledge Discovery, Modeling and Simulation. pp. 39–44. http://dx.doi.org/10.1145/2023568.2023579.

Tasci, B., Tasci, G., Ayyildiz, H., Kamath, A.P., Barua, P.D., Tuncer, T., Dogan, S., Ciaccio, E.J., Chakraborty, S., Acharya, U.R., 2024. Automated schizophrenia detection model using blood sample scattergram images and local binary pattern. Multimedia Tools Appl. 83 (14), 42735–42763. http://dx.doi.org/10.1007/s11042-023-16676-0.

Tennessee Tech University, 2024. Tennessee Tech University's Vocabulary. URL https://www.tntech.edu/library/pdf/Qualitative-quantitative-search-terms.pdf. (Last accessed: 09 March 2025).

The Content Authority, 2024. The Content Authority - Qualitative Research Words. URL https://thecontentauthority.com/blog/words-related-to-qualitative-research. (Last accessed: 09 March 2025).

Van Eck, N.J., Waltman, L., 2017. Citation-based clustering of publications using CitNetExplorer and VOSviewer. Scientometrics 111, 1053–1070. http://dx.doi.org/10.1007/s11192-017-2300-7.

Waltman, L., Van Eck, N.J., 2012. A new methodology for constructing a publication-level classification system of science. J. Am. Soc. Inf. Sci. Technol. 63 (12), 2378–2392. http://dx.doi.org/10.1002/asi.22748.

Williams, R.T., 2024. Paradigm shifts: exploring AI's influence on qualitative inquiry and analysis. Front. Res. Metrics Anal. 9, 1331589.

Wolff, B., Seidlmayer, E., Förstner, K.U., 2024. Enriched BERT embeddings for scholarly publication classification. In: International Workshop on Natural Scientific Language Processing and Research Knowledge Graphs. Springer, pp. 234–243. http://dx.doi.org/10.1007/978-3-031-65794-8_16.

Wolfswinkel, J.F., Furtmueller, E., Wilderom, C.P., 2013. Using grounded theory as a method for rigorously reviewing literature. Eur. J. Inf. Syst. 22 (1), 45–55. http://dx.doi.org/10.1057/ejis.2011.51.

Yohan, P., Sasidhar, B., Basha, S.A.H., Govardhan, A., 2014. Automatic named entity identification and classification using heuristic based approach for telugu. Int. J. Comput. Sci. Issues (IJCSI) 11 (1), 173.

Zhang, T., Shi, J., Situ, L., 2021. The correlation between author-editorial cooperation and the author's publications in journals. J. Inf. 15 (1), 101123. http://dx.doi.org/10.1016/j.joi.2020.101123.

Zhang, L., Sun, B., Shu, F., Huang, Y., 2022. Comparing paper level classifications across different methods and systems: an investigation of nature publications. Scientometrics 127 (12), 7633–7651. http://dx.doi.org/10.1007/s11192-022-04352-3.