RESEARCH ARTICLE

AI for AI: Using AI methods for classifying AI science documents

Evi Sachini[1], Konstantinos Sioumalas-Christodoulou[1,2], Stefanos Christopoulos[1,3], and Nikolaos Karampekios[1]

[1]National Documentation Centre (EKT), Palaio Faliro, Greece
[2]Department of History and Philosophy of Science, National and Kapodistrian University of Athens, Athens, Greece
[3]Cadence Design Systems, 85622 Munich, Germany

Keywords: article-level analysis, artificial intelligence, classification, neural networks, science

ABSTRACT

Subject area classification is an important first phase in the entire process involved in bibliometrics. In this paper, we explore the possibility of using automated algorithms for classifying scientific papers related to Artificial Intelligence at the document level. The current process is semimanual and journal based, a realization that, we argue, opens up the potential for inaccuracies. To counter this, our proposed automated approach makes use of neural networks, specifically BERT. The classification accuracy of our model reaches 96.5%. In addition, the model was used for further classifying documents from 26 different subject areas from the Scopus database. Our findings indicate that a significant subset of existing Computer Science, Decision Science, and Mathematics publications could potentially be classified as AI-related. The same holds in particular cases in other science fields such as Medicine and Psychology or Arts and Humanities. The above indicate that in subject area classification processes, there is room for automatic approaches to be utilized in a complementary manner with traditional manual procedures.

1. INTRODUCTION

In this paper, we explore the possibility of using automated algorithms for classifying documents that are used for bibliometric purposes. Rather than opting for a semimanual classification process, we opt for a fully automated approach by making use of a pretrained BERT model. The paper's classification subject matter is a subfield of Computer Science, that of "Artificial Intelligence" (AI).

The structure is as follows. In this section, we discuss the classification of science outputs, current practices by bibliometric databases, and the limitations of existing approaches. We continue therein and proceed to address the growing practice of bibliometric classification based on algorithmic approaches. The extended nature of this section is based on the multiple points of entry we wish to address to best frame the building blocks of the paper. Section 2 posits the subject matter—that is, the empirical case study of the paper. Then, the research question and (in Section 3) the objective are communicated. In Section 4, the contribution of this work follows. In Section 5, the methodology the data set and its (pre)processing, and the utilized model are addressed. The results are presented in Section 6 and the paper comes to an end with a discussion.

Here, we discuss the classification of science outputs, the process that is being conducted, and the problems associated with it. Also, we describe the rise of a number of automated

an open access journal

Citation: Sachini, E., Sioumalas-Christodoulou, K., Christopoulos, S., & Karampekios, N. (2022). AI for AI: Using AI methods for classifying AI science documents. *Quantitative Science Studies*, *3*(4), 1119–1132. https://doi.org/10.1162/qss_a_00223

DOI: https://doi.org/10.1162/qss_a_00223

Peer Review: https://publons.com/publon/10.1162/qss_a_00223

Received: 12 February 2022
Accepted: 16 October 2022

Corresponding Author: Konstantinos Sioumalas-Christodoulou
ksioumalas@ekt.gr

Handling Editor: Ludo Waltman

The MIT Press

approaches that can help steer this classification process into the future. For many decades, the problem of science classification has attracted the attention of philosophers and scientists—indeed, this work can be traced back to the 19th century (Dolby, 1979; Vickery, 1958). The practice of classification is understood as a process of arranging things "in groups which are distinct from each other, and are separated by clearly determined lines of demarcation" (Durkheim & Mauss, 1963, p. 4). However, nature, and therefore science, with its inherent complexity, does not conform to a particular categorization or hierarchical structuring (Bryant, 1997). Hence, there is no singular or perfect classification (Glänzel & Schubert, 2003). However, despite such inherent limitations, the practicalities in terms of organizing, studying, and analyzing scientific knowledge through classifications are well recognized and, indeed, invaluable.

Science subject area classification is what we are interested in. This refers to the hierarchical structure(s) and method(s) followed for organizing and documenting scientific content (mostly relating to science outputs, such as scientific publications, conference proceedings, and books), employing a formal broad-to-narrow categorization principle. Such schemas are made use of in multiple cases, one of which is research monitoring and evaluation through bibliometric analysis. Major institutional sectors (Organisation for Economic Co-operation and Development [OECD]), organizations (National Science Foundation [NSF]) and bibliometric groups (Centre for Science and Technology Studies, Science-Metrix) have long applied such classes to the field of bibliometrics. The objective of this formalized arrangement is to stack and settle in a logical manner the growing scientific content (referring to the ever-expanding field of science as a human endeavor) and to monitor, compare, and evaluate research outputs across different scientific disciplines at multiple levels (e.g., sector, regional, or country). As such, understanding the process of constructing the conceptual and methodological schemas with respect to the subject area classification of scientific outputs is a key point in the entire process.

In view of this, many classification schemes of science outputs have been proposed, each of them entailing different granularity and hierarchy levels. Although each scheme has its own level of complexity and sophistication, certain criteria have been constructed to compare and evaluate them (Rafols & Leydesdorff, 2009).

Traditional bibliometric databases, offering paid services, have been developing their own methodology for the assessment and classification of the underlying subject areas of science outputs. For instance, "Broad" subject areas (titles) in Scopus are classified under the following four broad subject clusters: Life Sciences, Physical Sciences, Health Sciences, and Social Sciences & Humanities. These are then further divided into 27 major subject areas and 300+ minor subject areas—fields (Scopus, 2020; Scopus, 2021a).

Scopus has a clearly stated selection policy and a board of selection experts for undertaking the classification process. Accordingly, the subject areas for the classification of a journal are assigned by the Scopus Content Selection and Advisory Board (CSAB). This is conducted during the Title evaluation process. An international group of scientists, researchers, and librarians, who comprise the CSAB in proportion to the major scientific disciplines, has the task of reviewing new titles and their suggested subject code on a continuous basis. Once the requested new title and its subject code are approved by this expert group, the new title will be classified and entered in the Scopus database accordingly (Scopus, 2021b).

The Web of Science (WoS) classifies all the indexed journals into approximately 252 groups called "Subject Categories." Such subject categories pertain to broader areas of Science, Social Sciences, and Arts and Humanities. Creating the schema involves assigning each journal to

one or more subject categories. The classification uses a number of heuristics, and its rather general description is provided by Pudovkin and Garfield (2002). WoS classification is not explicitly hierarchical, even though some subject categories can be considered as part of other broader ones. In addition, WoS contains categories that are explicitly broad (labeled as "Multidisciplinary") in order to describe the content of journals that publish across one broad area or across the entire field of science (Web of Science, 2020).

However, because it is often a challenging task to assign a journal to only a single category, overlapping coverage of categories can occur. This may complicate analysis. The same applies for classification at the document level. Although monodisciplinary classification schemas are suitable for highlighting the objects of the study solely within the scientific area or subject domain involved, such approaches fail to capture the bigger picture when it comes to scientific areas that are characterized by multidisciplinarity, such as Artificial Intelligence. Attributing a single discipline to a scientific document has proven to be a very hard task. In that capacity, any classification framework that intellectually assigns papers in such a way should be deployed with additional care.

On the other hand, with regard to documents that are considered as prepublication versions (preprints), a variety of databases exist, arXiv being one of the most established of this specific category of scientific documents (Wang & Zhan, 2019). Arxiv is a free distribution service and an open-access archive of over 2,000,000 scholarly articles in the fields of physics, mathematics, computer science, quantitative biology, quantitative finance, statistics, electrical engineering, and systems science (arXiv, 2022a). Within this service, a small group of volunteers, "moderators," consisting of interested experts from around the world, cursorily prescan new submissions. This is conducted typically only at the level of title and abstract and scans for appropriateness to the primary subject area which is proposed at the document level (arXiv, 2022b).

Over the years, a number of other journal-centered classifications have been developed—most of them being hierarchical. The OECD adopts a hierarchical classification scheme in which Science and Social Science are separated into six major subject categories. Within each subject category there are several minor subject categories (OECD, 2021). The NSF uses a two-level system in which journals are classified into 14 broad fields and 144 lower level fields known as CHI, after Narin and Carpenter's company, Computer Horizons, Inc., which developed it in the 1970s (Archambault, Beauchesne, & Caruso, 2011). Science-Metrix uses a three-level classification that classifies journals into exclusive categories using both algorithmic methods and expert judgement (Archambault et al., 2011). Glänzel and Schubert (2003) developed the KU Leuven ECOOM journal classification. Gómez-Núñez, Vargas-Quesada et al. (2011) used reference analysis to reclassify the SCImago Journal and Country Ranks (SJR) journals into 27 areas and 308 subject categories.

The most prevalent and widely used classification of literature into disciplines is based on journals. This approach relies on a rather simplistic assumption that a discipline can be defined through journal subject categories (Carpenter & Narin, 1973; Narin, 1976; Narin, Pinski, & Gee, 1976). Such an approach is not surprising—journals often serve as anchors for individual research communities, and new journals may signify the formations of disciplines (Milojević, 2020).

Although journal-level classification is the gold standard for many bibliometric databases and institutional sectors, it suffers from a number of problems, such as inaccurate taxonomies, misclassifications, and inability to perform granular analysis. This has given rise to more recent classification efforts based on document (article)-focused solutions. For example, Klavans and

Boyack (2017) found journal-based taxonomies of science to be more inaccurate than document-based ones and therefore argued against their use. Waltman and van Eck (2012) proposed a classification methodology regarding publications based on citation relations. Other findings were reported in a recent study that carried out direct comparison of journal- and article-level classifications (Shu, Julien et al., 2019). This study concluded that journal-level classifications have the potential to misclassify almost half of the papers. Leydesdorff and Rafols (2009) and Rafols and Leydesdorff (2009) find that journal-level classifications underperform when compared to article-level classification in relation to microlevel analysis, despite the fact that the former might still be useful for (nonevaluative) macro level analysis.

1.1. An Automated Way Forward

The issue with accuracy by way of making journal-level classifications is bound to increase, as both the number of journals that publish papers from multiple research areas and the number of papers published in those journals will continue to grow in the foreseeable future (Gómez, Bordons et al., 1996; Wang & Waltman, 2016). This numerical increase must be seen in tandem with the issue of subjectivity; that is, the human-centered process of attributing each journal to a specific subject class, category, etc. This is based on a manual process where a class of experts signal their willingness to accept or not a specific token within a class. This appears to be the case with CSAB and other classification processes for other bibliometric databases. Although expert groups are a standardized manner to legitimize policy options within knowledge societies, this approach is not immune to faults. For example, biases in elicitation, anchoring, and adjustment, and overconfidence are a select few of these faults (O'Hagan, 2019; Wilson, Shipley, & Davatzes, 2020; Tempelaar, Rienties, & Nguyen, 2020). In our case, overreliance on expert groups opens up the possibility for potential inaccuracies and wider problems of accountability, inclusion, and comprehensiveness.

A possible solution to this problem would be an approach in which experts take into consideration machine learning model suggestions in terms of field classification. In this way, the subject area classification process would be enhanced by means of cooperation between a model's suggestion and reviewer's opinion to formulate a more comprehensive approach and address the problem in a consistent manner.

Within this context, exploring, understanding, and utilizing underlying patterns (concepts, ideas, topics, words, etc.) that signal the subject area of a document or journal for performing classification tasks is of vital importance. Reviewing the relevant literature upon the plausible ways of performing text and document classification, different methods have been distinguished. These pertain to rule-based and data-based (machine learning) methods. Rule-based methods tend to require substantial engineering effort and deep knowledge of the domain of interest (Minaee, Kalchbrenner et al., 2021). In contrast, machine learning methods require less engineering effort but higher computation power, as they rely solely on observation data. As a result, rule-based methods, while widely used in the past, are quickly being replaced by data-driven approaches.

In the machine learning domain, neural network-based methods have come into the spotlight in recent years as their performance substantially exceeds all previous techniques. In particular, Transformer models (Vaswani, Shazeer et al., 2017) have achieved state-of-the art results in multiple tasks, including text classification. This is achieved by applying a stacked encoder-decoder architecture with self-attention mechanisms which replace the recurrent layers commonly used in the past. Since then, a multitude of Transformer-based models have

been developed, the most popular being OpenAI's GPT models (Radford & Narasimhan, 2018) and Google's BERT (Devlin, Chang et al., 2019). The latter is widely used for text classification among other tasks (Lan, Chen et al., 2020; Liu, Ott et al., 2019).

As mentioned, BERT is a language representation model based on the original Transformer Architecture. What makes it stand out is the use of bidirectional self-attention layers which better mimic the way humans process text. It is trained with a two-step approach. First, the network is trained on unlabeled data over multiple training tasks and, then, further fine-tuned on different tasks using labeled data. On top of that, WordPiece embeddings are used as input to the network along with several special tokens. The architecture of BERT makes it amenable to be used as a base language representation layer that can be built upon without necessitating time- and resource-consuming training of the language model from scratch. This, along with the fact that there exist multiple variations of pretrained BERT networks, makes it an ideal network to be used for natural language processing tasks.

In this study, we opt for a fully automated process of document classification based on BERT. The paper's classification subject matter is a subfield of Computer Science, that of AI.

This field of AI is attracting ever-increasing interest from the research community and public policy-makers, as well as members of the private sector, due to its high potential for the transformation of a variety of human activities.

Although in this paper we will make use of BERT, for literature purposes, it is important to note that further to neural networks, there are other, less compute-intensive methods that provide sufficient performance. Chowdhury and Schoen (2020) identify various techniques for the task of classifying research papers into three different fields (namely, Science, Business, and Social Science). These algorithms include Support Vector Machines, Naive Bayes, Decision Trees, and clustering algorithms such as K-Nearest-Neighbor. Furthermore, keyword-based techniques are frequently used for text classification. Automatic keyword extraction is the process of identifying key terms, key phrases, key segments, or keywords from a document that can appropriately represent the subject of the document (Beliga, Meštrović, & Martinčić-Ipšić, 2015). These keywords act as inputs to machine learning algorithms for text classification tasks. Onan, Korukoglu, and Bulut (2016) used a series of different keyword extraction methods and combined various machine learning models and ensemble techniques to achieve an accuracy of about 90% on the ACM and Reuters-21578 data sets. However, such processes of finding the "optimal" algorithms that perform best at a specific task require significant engineering effort in terms of time, computation, and complexity. BERT, however, provides more flexibility, as it is conceptually simple and can be used for a wide range of tasks without substantial task-specific architecture modifications (Devlin et al., 2019). This realization adds empirical evidence for using BERT as a base language representation layer for NLP problems (González-Carvajal & Garrido-Merchán, 2020; Minaee et al., 2021).

As a final note before proceeding to the next section, we would like to mention that our model, as any data-driven approach, strictly depends on the data that it is being trained with, meaning that the model reflects the current state of what AI is perceived to be. As new techniques and methods are invented, the state of AI will be enriched. However, as long as there is no extreme deviation between the two states, the model's performance should not regress significantly. In any case, given the dynamic nature of the AI field, data-driven models should take into account advances on the research front (i.e., new conceptual schemas and refined definitions of subject areas). That is, models should be fine-tuned in accordance with the cutting-edge literature by including representative papers into the data set.

2. RESEARCH QUESTION

We seek to examine whether the current process of conducting bibliometric analysis, which focuses on subjective classification of subject areas, can be enhanced by way of making use of purely automated techniques in classifying these areas.

Specifically, we seek to provide an answer to the following question: Is it possible by using data science techniques to provide a comprehensive bibliometric grouping of a specific thematic area (i.e., that of AI) in an automated manner?

3. RESEARCH OBJECTIVE

Given the above research question, the objective of the paper is to build an automatic mechanism that classifies documents related to AI. This will be conducted by making use of AI techniques.

4. CONTRIBUTION

In attempting to address the research question, this paper contributes to a number of thematic areas. Firstly, concerning bibliometrics we seek to enhance the current process of conducting bibliometric analysis. By providing a classification process of subject areas based purely on automated techniques, this will contribute to the comprehensiveness of the categories constructed during the classification processes. That is, the subject area classification process can be enhanced by means of cooperation between a model's suggestion (such as ours) and reviewer's opinion prior to the subject area finalization.

This leads to the second contribution. Given that our classification process is conducted at the level of documents (articles), this approach preserves the ability to discern disciplinary differences that potentially exist among individual articles published in the same journal.

Third, in this paper we show that it is possible to use a neural network, pretrained on general vocabulary, as an embedding layer to undertake domain-specific text classification. In addition, we expand on the process of creating a data set while making our model available for public use.

5. METHODOLOGY AND DATA

To reach the aforementioned objective, we implement a number of methodological steps. In this section, we discuss the data set, specifically journal selection and data retrieval, preprocessing, and the model implemented.

5.1. AI Journal Selection and Data Retrieval

With the aim of collecting a set of exploratory variables for our target binary variable ("AI" or "non- AI" papers), we probed into the abstracts of each corresponding paper. We intentionally excluded keywords as a dependent variable for our analysis purposes because authors often choose to put keywords that will attract the readers' interest so as to correspond to the main search engine keywords (Vincent-Lamarre & Larivière, 2021). This, most of the time, is done to the detriment of the accuracy of the wording and or the specific subject matter of the publication. Precisely because the abstract is clearly more extensive and describes in more detail the objectives, results, and insights of the article, it is a more privileged "place" to leverage any related information with the subject of our study: AI.

With the aim of collecting data and labels wherein the classification process occurs at both journal and document level, two bibliometric databases were probed. Specifically, the Scopus Database was indexed for journals which included "Artificial Intelligence," "Machine Learning," "Neural Networks," and "Neural Computing" in their title. This methodological approach is anchored in recent research performed by Yamashita, Murakami et al. (2021) in which terms such as "Machine Learning" and "Neural Network" constitute the strongest key AI terms appearing with high frequency in many databases. In total, 49,584 abstracts were collected from these journals for the AI class of the data set. In addition to the Scopus database, the arXiv data set—in which the classification process occurs at document level—was used to extract further data for the AI label. Recent studies have also utilized arXiv data for bibliometric analysis purposes (Okamura, 2022). When probing the arXiv data set, all abstracts from cs.AI (Artificial Intelligence) and cs.LG (Machine Learning) were considered. Approximately 122,911 abstracts were collected from the arXiv data set. With the aim of including cutting-edge research in the knowledge domain, top conferences in the field, such as NeurIPS, ICML, and ICLR, are included in the arXiv data set (approximately 8,000 papers).

Concerning the non-AI class of the data set, the Scopus database was used for collecting the necessary abstracts. Specifically, for each of the 26 subject areas—excluding the Multidisciplinary subject area—2,000 abstracts for every year since and including 2012 were collected. Any journals classified as AI were omitted. In total, approximately 520,000 abstracts were collected.

We decided to proceed with a manual approach[1] for downloading more than half a million abstracts from Scopus. All papers with no abstract available were deleted. As Table 1 suggests, the AI class consists of 172,495 abstracts. Of these, 129,371 (75%) are used for training while the remaining 43,124 (25%) are used for validation. Of the non-AI abstracts, 249,552 (50%) are used for training and validation purposes with a split of 75% (187,164) and 25% (62,388) respectively. The remaining abstracts are put in the test set, which is used to extract statistics about the predictive capabilities of the model on the different subject areas.

5.2. Preprocessing

Text preprocessing is a fundamental part of the classification process as it applies transformations to the text with the goal of making the learning process smoother. In addition, when using a pretrained model, the text which serves as input to the network should mimic the format of the text used to train the model. Consequently, our text preprocessing mirrors that used in the original BERT paper (Wu, Schuster et al., 2016) which makes use of WordPiece tokenization. Text is converted to lowercase because we use an uncased BERT model. Finally, it is crucial that all copyright messages as well as LaTeX characters are removed from the abstracts, because the model takes advantage of them. Maintaining them may result in better validation accuracy but ultimately the model ends up misclassifying a lot of examples in the test set.

5.3. The Model

As already noted, we use a smaller configuration of BERT. Our model consists of six Transformer Blocks, has a hidden layer size of 768 and uses 12 Attention Heads ($L = 6$, $H = 768$, $A = 12$). On top of that, we use a fully connected layer, followed by binary cross-entropy

[1] Scopus has a rate limit of 10,000 requests on a weekly basis for abstract retrievals on its API service (https://dev.elsevier.com/api_key_settings.html). We proceeded with our analysis in terms of performing subsequent queries on Scopus website (using Scopus' advanced search option).

Table 1. Distribution of AI label and Non-AI label abstracts with respect to documents within Scopus and Arxiv databases.

Data set	Training set	Validation set	Test set
AI label	129,371	43,124	–
Non-AI label	187,164	62,388	249,552

loss. The AdamW optimizer ($\beta1$ = 0.9, $\beta2$ = 0.999) was used with an initial learning rate of 3e−5 which decays to 1e−6 and a warm-up period of 10% of the total training steps. A batch size of eight is used. We train for just two epochs, as our training data set is quite large. We chose this particular BERT configuration because we wanted the model to be large enough to achieve good performance after training for a small number of epochs while also being small enough to fit on a single GPU to make it more accessible.

6. RESULTS

6.1. Classification Results on the Validation Set

After training our model, we run inference on the validation set. Our model achieves a validation accuracy of 96.5% (= (TP + TN)/(TP + FP + TN + FN)—see Table 2). Considering as "ground truth" Scopus' and arXiv's classification schemas, the totality of misclassified scientific papers is 3,684 (3.5%). Table 2 provides evidence with relation to the confusion matrix of the papers included in the validation set. Specifically, taking into account the AI class, 3.05% (False Negative) of the papers were classified as non-AI-related although they are labeled as AI-related.

On the other hand, concerning the non-AI class, the vast majority of papers (96.20%) were classified accordingly and the remaining 3.80% were classified as AI-related. This particular category is of importance and its implications are further analyzed in the following part of the results' section.

6.2. Classification Results on the Test Set

The majority of the classification results on the test set are within expectations. However, 3.72% of the cases were classified as AI-related. Specifically, we find that even though our test set consists of papers from journals not classified as AI-related, our model classifies a considerable portion of them as AI-related. This discrepancy accounts for the model's conceptual perception in terms of what could be related to AI. In particular, this discrepancy may be due to several reasons.

Table 2. The confusion matrix in relation to the number of papers selected for the validation set, with the total number being equal to 105,512 scientific papers.

	Frequency (# of papers)	**Percentage**
True Positive (TP)	41,809	96.95
False Negative (FN)	1,315	3.05
True Negative (TN)	60,019	96.20
False Positive (FP)	2,369	3.80

The first may be due to the model itself, its learning capabilities, and the training procedure in comparison to any other potential model. Second, the data used to train the model may be a cause. Specifically, our training data set consists of an integration of data used for training our model. There are data coming from Scopus and arXiv that follow different classification schemes, so the model is expected to emulate the classification rationale behind both of these schemes. In addition, as we discuss below in more detail (Section 7), we expect that due to the known limitations of journal-level classification (Klavans & Boyack, 2017) there are possibly some mislabeled papers. That is, a journal labeled as non-AI related may have a small percentage of papers related to AI and vice versa. This introduces "noise" in the training set, which can affect model performance. Last, irrespective of the model's predictive capabilities, the fact that we are making predictions on the article level but the test set labels are assigned by means of journal-based classification can lead to a slight divergence between the two due to their inherent differences.

Despite the above, by reading through some of the abstracts of the test set classified as AI-related, we note that a substantial portion of them are indeed AI-related. We make available a relevant file with such abstracts marked (see the relevant GitHub link in the "Data Availability" section).

In any case, this does not imply that one classification schema is necessarily better than the other. Journal-based classification methods subject to expert opinion and document-based methods implemented through automatic processes should be utilized in a complementary manner.

In Table 3 we observe that for subject areas such as Computer Science (22.39%), Decision Sciences (16.17%) and Mathematics (10.41%), areas which are closely linked to the study of AI, the model classifies a significant portion of papers as related to AI—something that is not the case in the standard Scopus approach. By examining the corresponding subfields of the aforementioned subject areas as classified within the Scopus database, the above result stands to reason (Scopus, 2021c). Specifically, apart from Computer Science, which incorporates AI as a scientific subfield, Decision Sciences contains subfields such as Information Systems and Management, Statistics, and Probability and Uncertainty. In a similar manner, Mathematics includes fields such as Computational Mathematics, Modelling and Simulation, and Control and Optimization. This is in line with the relevant bibliography, as all the above scientific subfields often make use of AI techniques as a means to perform tasks and analysis relevant to the current object of study and vice versa[2] (Brodie & Mylopoulos, 2012; Ören & Zeigler, 1987; Shafer, 1987; Shin & Xu, 2017).

With regard to other subject areas such as those related to Medicine, Dentistry, or Veterinary Science, the portion of the AI related papers is lower (below 1%). A potential reason could be that the subfields under the label of Medicine (0.79%) are associated mainly with "basic research" studies (Hematology, Dermatology, Embryology, etc.) rather than using applied techniques and methodological/algorithmic approaches and concepts such as AI. The same holds for the subject areas of Dentistry (Dental Hygiene, Orthodontics, etc.) and Veterinary Science (Equine, Small Animals, etc.) with 0.30% and 0.14%.

The cases of Psychology and Arts and Humanities present particular interest in as much as AI is being explored as a social and technological phenomenon or a classification and

[2] However, in our case, it seems that the "direct" should hold. That is, taking as given the Scopus classification, such articles should utilize AI techniques to fulfill objectives within the borders of their specialism.

Table 3. Results pertaining to the percentage of documents identified as AI-related through our approach within the 26 different subject areas.

Subject areas	AI-related (%)
Computer Science	22.39
Decision Science	16.17
Mathematics	10.41
Engineering	5.80
Health Professions	4.85
Neuroscience	3.81
Arts and Humanities	3.38
Business, Management, Accounting	3.21
Energy	2.82
Earth and Planetary Sciences	2.77
Social Sciences	2.71
Chemistry	2.39
Physics and Astronomy	2.29
Psychology	1.68
Economics and Econometrics	1.64
Environmental Sciences	1.51
Chemical Engineering	1.46
Biochemistry, Genetics, Molecular Biology	1.41
Pharmacology, Toxicology, Pharmaceutics	1.18
Material Science	0.92
Agricultural and Biological Sciences	0.85
Medicine	0.79
Immunology and Microbiology	0.56
Nursing	0.45
Dentistry	0.30
Veterinary Science	0.14

forecasting tool. For example, probing into our data set and examining the abstracts of such areas, certain studies explore the role and the impact of AI on society (Arts and Humanities) or examine the construction and evaluation of ML binary classification models (Arts and Humanities). Others utilize data science techniques to improve the classification rate of guilty and innocent subjects (Psychology) or apply ML models, such as Random Forest and Lasso Regression, to better predict future suicidal behavior (Psychology).

The above findings indicate that scientific publications classified as belonging to existing scientific fields should potentially be classified as "AI-related" as well. At the very least, examiners should heed the model's suggestions and take the necessary steps to confirm or refute them.

7. DISCUSSION

Analyzing, managing, and extrapolating information from an ever-increasing pool of data seems to be a problem of the future (including the present). Although not the case, classification and taxonomies of the collected data should be considered to be part and parcel of the same process. Herein, we focused on bibliometric data and showed that it would be beneficial for experts to take into account classification suggestions made by Natural Language Processing (NLP) models to increase the comprehensiveness of current classification schemas.

This appears to be necessary, as the current bibliometric classification process is manual with a small pool of experts classifying documents in a relatively subjective manner based on their expert knowledge. Although expert groups are a standardized way to legitimize policy options within knowledge societies, this is an approach not immune to faults. For example, biases in elicitation, anchoring, and adjustment, as well as overconfidence, are a select few of these faults (O'Hagan, 2019; Tempelaar et al., 2020; Wilson et al., 2020). This opens up the possibility for misclassifications and wider problems of accountability and inclusion. In our case, a small but not inconsiderable proportion of scientific documents (3.7%) was classified as related to AI although not indicated as such within traditional bibliometric databases. In addition, with regard to the analysis that pertains to the classification results of the test set, findings suggest that a portion of scientific documents that ranges from 0.14% (Veterinary Science) to 22.39% (Computer Science) was classified as AI-related throughout 26 different subject areas.

This paper contributes in a number of ways, be they bibliometrics or AI focusing on neural networks. Concerning bibliometrics, this paper has provided ways to enhance the current process of conducting bibliometric analysis, which focuses on subjective classification of subject areas. By making use of purely automated techniques, based on existing data emanating from experts' classification schemas, we propose a more inclusive approach that can be utilized and generalized in subsequent studies with similar objectives (i.e., patent classification). Importantly, this approach can be adopted in a complementary manner to the current classification process based on experts' suggestions. A potential avenue of this is for the responsible parties (experts, preassigned boards, etc.), during the classification process, to effortlessly take into consideration the "automatic" suggestions in terms of subject area classification and at the second stage come to their own conclusion. A second contribution concerns the level at which this endeavor was pursued. Rather than focusing on the journal level, we zoomed into the level of scientific documents.

Concerning AI, and more specifically neural networks, we capitalized upon the existing knowledge on using neural network models and explored whether it is possible to achieve sufficiently high accuracy on classifying research papers by using a relatively small version of BERT pretrained on nondomain-specific data. In addition, we expand on the process of creating a data set and we make our model publicly available.

As in every study, limitations pertain to the time element and the data set. The concept of AI is a dynamic one. As time goes on and computational power increases, the prevalent AI techniques evolve. For example, from the 1990s to 2000s, rule-based techniques were prevalent, whereas nowadays we are witnessing the neural network era. Thus, any attempt to

comprehensively and exhaustively denote the full array of its denotation should take into consideration such a realization. Hence, finding a common, "static" ground truth proves to be quite a hard undertaking.

Second, this analysis depends on the data acquired. Our finalized data set consists of an integration of data in terms of two conceptually different classification methodologies: a journal-based classification (data collected from the Scopus database) and a document-based classification (data collected from the arXiv database). As regards arXiv's case, the subject area classification is based on expert opinion/suggestion and occurs at the document level, thus meaning the subject areas attributed to each paper are considered to be largely accurate. However, this does not imply that the classification process is exhaustive. This should be taken to mean that while each paper may be accurately attributed to a specific subject area, it may well be the case that the specific paper can additionally be classified in other subject areas. Indeed, this has been shown to be the case not only within the context of this paper but more largely within the field of interdisciplinary sciences. This bears the issue of multilabel classification—a paper can be attributed to more than one subject area.

Taking into consideration the above realizations in conjunction with the strict data procurement process we followed, the collected AI-related science documents are labeled as accurately as possible given the circumstances. That is in contrast to documents collected for the non-AI label. As discussed, most bibliometric databases use a journal-based classification scheme, with the inherent problems this entails. This leads to our data set being somewhat "noisy."

However, Deep Neural Networks do not just memorize data. They tend to first learn the simple patterns that are common to multiple training examples. Only after multiple training epochs do they begin memorizing the noise in the data, which leads to the phenomenon known as "overfitting." They are quite robust to extreme amounts of label noise provided that the training data are ample and that the training process is stopped early so as to avoid overfitting (Rolnick, Veit et al., 2017). Thus, we believe that training with data labeled in a journal-based manner does not significantly affect model performance.

This ambiguity created from the fluid definition of AI and the small subset of incorrect labels in our data set poses a challenge in how to accurately quantify performance. Indeed, accuracy metrics in such cases have a degree of uncertainty, but taking into account the above points, we believe that the error margins of such metrics are quite muted.

Concerning future research avenues, analysis of other scientific subject areas stands as an option. One potential avenue would be to extend this approach to a multiclass classification schema by way of attributing to a single document multiple scientific areas or categories. In addition, classification of technology outputs is a promising road. Automated processes can be exploited for the classification of, for example, patents and industrial designs and the validation of the accepted classification taxonomies.

AUTHOR CONTRIBUTIONS

Evi Sachini: Conceptualization, Resources. Konstantinos Sioumalas-Christodoulou: Conceptualization, Formal analysis, Methodology, Supervision, Validation, Writing—Original draft, Writing—Review & editing. Stefanos Christopoulos: Conceptualization, Data curation, Investigation, Software, Visualization, Writing—Original draft, Writing—Review & editing. Nikolaos Karampekios: Conceptualization, Methodology, Project administration, Supervision, Writing—Original draft, Writing—Review & editing.

COMPETING INTERESTS

The authors have no competing interests.

FUNDING INFORMATION

This research received no specific grant from any funding agency (institutional, private and/or corporate financial support).

DATA AVAILABILITY

Scopus' legal constraints with respect to data sharing allow the authors to share up to 2,000 abstracts. The aforementioned data and code used in this paper can be found in Zenodo (Sachini, Sioumalas-Christodoulou et al., 2022).

REFERENCES

Archambault, É., Beauchesne, O. H., & Caruso, J. (2011). Towards a multilingual, comprehensive and open scientific journal ontology. In B. Noyons, P. Ngulube, & J. Leta (Eds.), *Proceedings of the 13th International Conference of the International Society for Scientometrics and Informetrics (ISSI)* (pp. 66–77). Durban, South Africa.

arXiv. (2022a). Available at: https://arxiv.org/

arXiv. (2022b). *Moderators*. Available at: https://arxiv.org/moderators/

Beliga, S., Meštrović, A., & Martinčić-Ipšić, S. (2015). An overview of graph-based keyword extraction methods and approaches. *Journal of Information and Organizational Sciences*, *39*(1). https://jios.foi.hr/index.php/jios/article/view/938

Brodie, M. L., & Mylopoulos, J. (Eds.). (2012). *On knowledge base management systems: Integrating artificial intelligence and database technologies*. Springer Science & Business Media. https://doi.org/10.1007/978-1-4612-4980-1

Bryant, R. E. (1997). Discovery and decision: Exploring the metaphysics and epistemology of scientific classification. *Doctoral dissertation*, University of Edinburgh.

Carpenter, M. P., & Narin, F. (1973). Clustering of scientific journals. *Journal of the American Society for Information Science*, *24*(6), 425–436. https://doi.org/10.1002/asi.4630240604

Chowdhury, S., & Schoen, M. P. (2020). Research paper classification using supervised machine learning techniques. *2020 Intermountain Engineering, Technology and Computing (IETC)* (pp. 1–6). https://doi.org/10.1109/IETC47856.2020.9249211

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805*. https://doi.org/10.48550/arXiv.1810.04805

Dolby, R. G. (1979). Classification of the sciences: The nineteenth century tradition. In R. F. Ellen & D. Reason (Eds.), *Classifications in their social context*. London: Academic Press.

Durkheim, E., & Mauss, M. (1963). *Primitive classification*. Chicago: University of Chicago Press.

Glänzel, W., & Schubert, A. (2003). A new classification scheme of science fields and subfields designed for scientometric evaluation purposes. *Scientometrics*, *56*(3), 357–367. https://doi.org/10.1023/A:1022378804087

Gómez, I., Bordons, M., Fernandez, M., & Méndez, A. (1996). Coping with the problem of subject classification diversity. *Scientometrics*, *35*(2), 223–235. https://doi.org/10.1007/BF02018480

Gómez-Núñez, A. J., Vargas-Quesada, B., de Moya-Anegón, F., & Glänzel, W. (2011). Improving SCImago Journal & Country Rank (SJR) subject classification through reference analysis. *Scientometrics*, *89*, 741. https://doi.org/10.1007/s11192-011-0485-8

González-Carvajal, S., & Garrido-Merchán, E. C. (2020). Comparing BERT against traditional machine learning text classification. *arXiv:2005.13012*. https://doi.org/10.48550/arXiv.2005.13012

Klavans, R., & Boyack, K. W. (2017). Which type of citation analysis generates the most accurate taxonomy of scientific and technical knowledge? *Journal of the Association for Information Science and Technology*, *68*(4), 984–998. https://doi.org/10.1002/asi.23734

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). ALBERT: A lite BERT for self-supervised learning of language representations. *arXiv:1909.11942*. https://doi.org/10.48550/arXiv.1909.11942

Leydesdorff, L., & Rafols, I. (2009). A global map of science based on the ISI subject categories. *Journal of the American Society for Information Science and Technology*, *60*(2), 348–362. https://doi.org/10.1002/asi.20967

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., ... Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv:1907.11692*. https://doi.org/10.48550/arXiv.1907.11692

Milojević, S. (2020). Practical method to reclassify Web of Science articles into unique subject categories and broad disciplines. *Quantitative Science Studies*, *1*(1), 183–206. https://doi.org/10.1162/qss_a_00014

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2021). Deep learning-based text classification: A comprehensive review. *ACM Computing Surveys*, *54*(3), 1–40. https://doi.org/10.1145/3439726

Narin, F. (1976). *Evaluative bibliometrics: The use of publication and citation analysis in the evaluation of scientific activity*. Cherry Hill, NJ: Computer Horizons.

Narin, F., Pinski, G., & Gee, H. H. (1976). Structure of the biomedical literature. *Journal of the American Society for Information Science*, *27*(1), 25–45. https://doi.org/10.1002/asi.4630270104

OECD. (2021). *OECD category scheme*. https://help.prod-incites.com/inCites2Live/filterValuesGroup/researchAreaSchema/oecdCategoryScheme.html

O'Hagan, A. (2019). Expert knowledge elicitation: Subjective but scientific. *The American Statistician*, *73*(sup1), 69–81. https://doi.org/10.1080/00031305.2018.1518265

Okamura, K. (2022). Scientometric engineering: Exploring citation dynamics via arXiv eprints. *Quantitative Science Studies*, *3*(1), 122–146. https://doi.org/10.1162/qss_a_00174

Onan, A., Korukoglu, S., & Bulut, H. (2016). Ensemble of keyword extraction methods and classifiers in text classification. *Expert Systems with Applications*, *57*, 232–247. https://doi.org/10.1016/j.eswa.2016.03.045

Ören, T. I., & Zeigler, B. P. (1987). Artificial intelligence in modelling and simulation: Directions to explore. *Simulation*, *48*(4), 131–134. https://doi.org/10.1177/003754978704800403

Pudovkin, A. I., & Garfield, E. (2002). Algorithmic procedure for finding semantically related journals. *Journal of the American Society for Information Science and Technology*, *53*(13), 1113–1119. https://doi.org/10.1002/asi.10153

Radford, A., & Narasimhan, K. (2018). *Improving language understanding by generative pre-training*. training. https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf

Rafols, I., & Leydesdorff, L. (2009). Content-based and algorithmic classifications of journals: Perspectives on the dynamics of scientific communication and indexer effects. *Journal of the American Society for Information Science and Technology*, *60*(9), 1823–1835. https://doi.org/10.1002/asi.21086

Rolnick, D., Veit, A., Belongie, S., & Shavit, N. (2017). Deep learning is robust to massive label noise. *arXiv:1705.10694*. https://doi.org/10.48550/arXiv.1705.10694

Sachini, E., Sioumalas-Christodoulou, K., Christopoulos, S., & Karampekios, N. (2022). Model and code of the scientific paper "AI for AI: Using AI methods for classifying AI science documents." *Zenodo*. https://doi.org/10.5281/zenodo.7223811

Scopus. (2020). *Content coverage guide*. https://www.elsevier.com/__data/assets/pdf_file/0007/69451/Scopus_ContentCoverage_Guide_WEB.pdf

Scopus. (2021a). *What is the complete list of Scopus Subject Areas and All Science Journal Classification Codes (ASJC)?* https://service.elsevier.com/app/answers/detail/a_id/15181/supporthub/scopus/

Scopus. (2021b). *Content policy and selection*. https://www.elsevier.com/solutions/scopus/how-scopus-works/content/content-policy-and-selection

Scopus. (2021c). *What are the most used subject area categories and classifications in Scopus?* https://service.elsevier.com/app/answers/detail/a_id/14882/supporthub/scopus/~/what-are-the-most-frequent-subject-area-categories-and-classifications-used-in/

Shafer, G. (1987). Probability judgment in artificial intelligence and expert systems. *Statistical Science*, *2*(1), 3–16. https://doi.org/10.1214/ss/1177013426

Shin, Y. C., & Xu, C. (2017). *Intelligent systems: Modeling, optimization, and control*. CRC Press. https://doi.org/10.1201/9781420051773

Shu, F., Julien, C. A., Zhang, L., Qiu, J., Zhang, J., & Larivière, V. (2019). Comparing journal and paper level classifications of science. *Journal of Informetrics*, *13*(1), 202–225. https://doi.org/10.1016/j.joi.2018.12.005

Tempelaar, D., Rienties, B., & Nguyen, Q. (2020). Subjective data, objective data and the role of bias in predictive modelling: Lessons from a dispositional learning analytics application. *PLOS ONE*, *15*(6), e0233977. https://doi.org/10.1371/journal.pone.0233977, PubMed: 32530954

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., ... Polosukhin, I. (2017). Attention is all you need. *arXiv:1706.03762*. https://doi.org/10.48550/arXiv.1706.03762

Vickery, B. C. (1958). *Classification and indexing in science*. Butterworths Scientific Publications.

Vincent-Lamarre, P., & Larivière, V. (2021). Textual analysis of artificial intelligence manuscripts reveals features associated with peer review outcome. *Quantitative Science Studies*, *2*(2), 662–677. https://doi.org/10.1162/qss_a_00125

Waltman, L., & Van Eck, N. J. (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology*, *63*(12), 2378–2392. https://doi.org/10.1002/asi.22748

Wang, L., & Zhan, Y. (2019). A conceptual peer review model for arXiv and other preprint databases. *Learned Publishing*, *32*(3), 213–219. https://doi.org/10.1002/leap.1229

Wang, Q., & Waltman, L. (2016). Large-scale analysis of the accuracy of the journal classification systems of Web of Science and Scopus. *Journal of Informetrics*, *10*(2), 347–364. https://doi.org/10.1016/j.joi.2016.02.003

Web of Science. (2020). *Research area schemas*. https://incites.help.clarivate.com/Content/Indicators-Handbook/ih-research-area-schemas.htm

Wilson, C. G., Shipley, T. F., & Davatzes, A. K. (2020). Evidence of vulnerability to decision bias in expert field scientists. *Applied Cognitive Psychology*, *34*(5), 1217–1223. https://doi.org/10.1002/acp.3677

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., ... & Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv:1609.08144*. https://doi.org/10.48550/arXiv.1609.08144

Yamashita, I., Murakami, A., Cairns, S., & Galindo-Rueda, F. (2021). Measuring the AI content of government-funded R&D projects: A proof of concept for the OECD Fundstat initiative. *OECD Science, Technology and Industry Working Papers*, No. 2021/09. Paris: OECD Publishing. https://doi.org/10.1787/7b43b038-en