IDENTIFYING THE DEVELOPMENT AND APPLICATION OF ARTIFICIAL INTELLIGENCE IN SCIENTIFIC TEXT*

James Dunham  
Center for Security and Emerging Technology  
Georgetown University  
james.dunham@georgetown.edu

Jennifer Melot  
Center for Security and Emerging Technology  
Georgetown University  
jennifer.melot@georgetown.edu

Dewey Murdick  
Center for Security and Emerging Technology  
Georgetown University  
dewey.murdick@georgetown.edu

February 18, 2020

ABSTRACT

We describe a strategy for identifying the universe of research publications relating to the application and development of artificial intelligence. The approach leverages arXiv’s corpus of scientific preprints, in which authors choose subject tags for their papers from a set defined by editors. We compose from these subjects a functional definition of AI-relevance with intuitive components, by learning the subject definitions from paper metadata, and then inferring the arXiv-subject labels of papers in Web of Science. We find predictive classification $F_1$ scores between .59 and .86 for AI-relevant subject models. For an all-subjects model, we see precision of .83 and recall of .85. We evaluate the out-of-domain performance of our classifiers against other sources of subject information and results from other methods. We find that for the high-level fields of study represented on arXiv, a supervised solution can generalize for inference in other corpora. This offers a method for identifying AI-relevant publications that updates at the pace of research output, without reliance on subject-matter experts for query development or labeling.

1 Overview

Study of the applications and development of artificial intelligence faces a definitional problem: AI is a moving conceptual target, understood differently across researchers and observers of the field [12]. This presents a challenge for analysts and policy-makers [25]. The proliferation of reports on AI describe only partially overlapping domains [1][18][2], suggesting that the delineation of the field affects the reliability of conclusions [26]. We describe a strategy for addressing this and identifying a universe of AI-relevant scientific publications for use in bibliometric work.

The approach relies on the success of Cornell’s arXiv project in attracting open-access preprints from subfields of computer science, physics, statistics, and other quantitative fields. Authors identify the subjects of papers they upload from a set of labels defined by editors. There are 39 subjects in computer science, including those we will consider relevant to AI: Artificial Intelligence, Computer Vision, Computation and Language (Natural Language Processing), Machine Learning, Multiagent Learning, and Robotics. These and other arXiv labels offer a particular ground truth

*We thank Kevin Boyack, Daniel Chou, Teddy Collins, Dick Klavans, and Ilya Rahkovsky for their feedback and ideas on this work. We are grateful to the team at Elsevier for extended discussions about the methodological details of a related project, and sharing expert-curated keywords and labeled data. Zihe Yang led the replication of the Elsevier approach to identifying AI-relevant research. Neha Tiwari contributed the descriptive analysis of arXiv and conference-paper data, and assisted with model development. For replication materials, see https://github.com/georgetown-cset/ai-relevant-papers

[https://arxiv.org](https://arxiv.org)
defined by the participation of an expert community. Additionally, arXiv’s implicit definition of subjects has the highly desirable characteristic of updating in real time, as opposed to less-favorable approaches that rely on keyword curation or annotation by subject-matter experts. Those alternatives tend to require maintenance over time, and as we demonstrate, a query that subject-matter experts calibrate to retrieve AI-relevant publications in 2019 may struggle to surface those from 2010.

We are keenly aware that the subjects comprising AI research and applications are contestable. Rather than argue for a single delineation, we offer an approach which requires only that an operational definition is composable from the subjects available to arXiv authors. The sensitivity of all subsequent analysis to that choice of relevant subjects can be assessed through ablation. Researchers may also add or remove particular subjects as appropriate for their analyses.

We implement this approach by training SciBERT classifiers on arXiv metadata and subject labels. Using the arXiv-trained models, we infer the subject relevance of papers in other corpora. The premise of identifying AI-relevant publications in this way is that a model trained on arXiv data will successfully generalize to other sets of publication data, which may significantly differ in content and subject distribution. This approach seems plausible when leveraging SciBERT’s pre-training, but the risk of overfitting to arXiv and gaps in its coverage are concerns we address below with a series of results.

First, to assess performance within arXiv, we partition the data and evaluate our models on a test set. We observe $F_1$ scores between .59 and .86 for one-versus-all subject models, and .84 for a model trained on labels collapsed to indicate AI-relevance for papers in any of these subjects. For comparison, we also assess a keyword-query solution and a keyword-learner hybrid developed for a recent bibliometric analysis of AI-relevant publications in Scopus [1]. Evaluation against arXiv labels yields $F_1$ scores of .55 and .66, respectively, for these methods.

We then report results from applying the models to Web of Science (WoS) publications. In the absence of ground-truth arXiv labels throughout WoS, we assess out-of-domain performance using other sources of topical information. We demonstrate rates of predicted subject relevance in selected journals, conference proceedings, and WoS categories. We find that for the high-level fields of study represented on arXiv, generalizing for inference in other corpora is feasible. This offers a method for identifying AI-relevant publications that updates at the pace of research output, without reliance on subject-matter experts for query development or labeling.

2 Development and applications of artificial intelligence

Scientific text offers insight into the development of a field: its analysis can identify the organization of research communities; their breakthroughs or stagnation; and progress from basic research to applications [e.g., 23] [5]. The obstacles to such inference are delineation of that field and the identification of emergent topics or technologies within it [7]. In reference to biotech and nanotech in prior decades, Mogoutov and Kahane write, “Their content and dynamic are difficult to track at a time when they are struggling to define what they are, what they include and exclude, and how they organize and classify themselves internally” [16]. A related problem is identifying as-yet-unknown topics within a field, without the benefit of historical perspective. Even in emergent areas, the distinction between “legacy technologies” and “emerging technology” may be incremental [10].

Recent analyses of AI research using query-based methods to delineate the field [17] [15] [19] have encountered these obstacles. Grappling with the problem of query development in bibliometric work on nanotechnology resulted in principled methods for term curation and their evaluation [16] [9] [14], from which studies of AI could benefit. Drawing from this literature, for example, Huang et al. develop a method for retrieving “big data” research that expands from an initial set of terms across iterations of discovery, manual review, expert checks, and tuning for performance [11].

Other approaches to delineation depend on or begin with the identification of relevant journals [8] or conferences [13] [21]. While appropriate for some analytic purposes, this method risks omitting relevant research in more general-audience venues or other disciplines, which may be a particularly acute problem for AI.

In review of the variety of methods for delineating the field of AI-relevant research, we note that beyond the methodological difficulties, the criteria for a system’s intelligence vary by observer and over time. In the typology developed

---

1 See https://arxiv.org/category_taxonomy
2 See https://clarivate.com/webofsciencegroup
3 See Dimensions, Scopus, or Semantic Scholar
4 We expect our approach to be similarly adaptable to scientific text in other collections, like Dimensions, Scopus, or Semantic Scholar, but we restrict this paper to WoS.
5 For a discussion of precisely what constitutes emerging technology, see [24].
by Russell and Norvig [20], definitions may emphasize behavior or reasoning, and evaluate it against human or rational standards. In recent survey research [12], AI researchers tended to prefer definitions that emphasized the correctness of decisions and actions, but often disagreed on what satisfied these requirements.

Our own interest in connecting policymakers to high-quality analysis of AI and its security implications requires an AI definition that is robust over time and covers both research and applications. As AI methods, tasks, and applications increase in diversity, expert-informed queries become increasingly impractical. The solution discussed within this paper is most applicable to the AI research community’s output, but the general approach applies to a wide variety of dynamic classifier-based definitions across many types of textual source materials. We therefore require criteria for identifying publications describing AI research or its applications that embrace variation in definitions across contexts and time.

3 Data

arXiv is organized into high-level domain repositories for physics, biology, computer science, statistics, and so forth. Each of these repositories further defines a set of subjects to organize its content. Authors select one or more subjects to describe each paper they submit. Editors later review these subject tags [6]. arXiv’s Computing Research Repository (CoRR) defines 39 subjects including artificial intelligence and machine learning.

We focus in this paper on six subjects that CoRR editors describe as related to AI: Artificial Intelligence, Computation and Language (NLP), Computer Vision and Pattern Recognition (CV), Machine Learning, Multiagent Systems, and Robotics. According to CoRR documentation, the Artificial Intelligence subject “[c]overs all areas of AI except Vision, Robotics, Machine Learning, Multiagent Systems, and Computation and Language (Natural Language Processing),” because these areas have their own subjects. It specifically “includes Expert Systems, Theorem Proving […], Knowledge Representation, Planning, and Uncertainty in AI.” The Machine Learning subject “[c]overs all aspects of machine learning research [and] is also an appropriate primary category for applications of machine learning methods.” Because these applications may have their own subject areas, CoRR documentation specifies, “If the primary domain of the application is available as another category in arXiv and readers of that category would be the main audience, that category should be primary.” Some explicit examples of this are papers on CV, NLP, information retrieval, speech recognition, and neural networks.

Using arXiv submissions in these categories as training data for subject classifiers, and defining AI-relevant research as the union of their positive predictions, is a useful framework for future researchers who may have differing needs or views on what constitutes AI. Adding Neural and Evolutionary Computing or Information Retrieval papers might be warranted in future work. We exclude them here for consistency with the CoRR editors’ description of the Artificial Intelligence subject, but in practice, we suggest evaluating how sensitive quantities of interest are to these choices.

The compositional effect of including or excluding some subjects will be modest due to patterns of cross-posting papers across related subjects. There are 3,464 papers in our data with Information Retrieval as their primary subject, and 42% also appear in one or more of the six subjects we consider AI-relevant here. Of the 2,942 papers with the primary category of Neural and Evolutionary Computing, 39% are cross-posted to at least one of our AI-relevant subjects, primarily Machine Learning.

From 2010 through 2019, authors submitted 1,060,321 papers to arXiv. The largest repositories at the end of this decade, counting by papers’ primary subjects, are physics (540,692), math (270,244), and computer science (194,627). Table 1 shows paper counts in the six computer science subjects we consider relevant. There are 85,670 whose primary subject, the first selected by authors, is one of these six. Authors can cross-post their papers under additional subjects, however, and when including these cross-posts there are 107,380 papers across the relevant subjects.

Our target for inference is the Web of Science (WoS) Core Collection. Training on arXiv is appealing for reasons we have described, but we ultimately care about performance in WoS or other more general knowledge bases, and many

---

6 The Center for Security and Emerging Technology (CSET) studies the security impacts of emerging technologies and delivers nonpartisan analysis to the policy community. See examples of reports that are dependent on various AI definitions at [https://cset.georgetown.edu/reports](https://cset.georgetown.edu/reports).

7 See [https://arxiv.org/category_taxonomy](https://arxiv.org/category_taxonomy).

8 We include machine learning papers from the statistics repository (stat.ML) in this subject. Cross-posting between the two categories is automatic. [https://arxiv.org/corr/subjectclasses](https://arxiv.org/corr/subjectclasses).

9 We restrict this effort to the last decade of arXiv papers to ensure reasonable numbers of papers in each subject in every year.
Table 1: arXiv contains 85,670 papers from 2010–2019 whose primary subject is one of the six we selected as relevant. 107,380 papers, or an additional 21,710, appeared in at least one of the six subjects. This includes cross-posts from other subjects.

| Subject                          | Papers with Primary Subject | Papers Including Cross-posts |
|----------------------------------|----------------------------|-----------------------------|
| Artificial Intelligence (cs.AI)  | 8,941                      | 19,964                      |
| Natural Language Processing (cs.CL) | 11,881                    | 15,361                      |
| Computer Vision (cs.CV)          | 28,309                     | 35,254                      |
| Machine Learning (cs.LG, stat.ML) | 30,175                     | 52,909                      |
| Multiagent Systems (cs.MA)       | 985                        | 2,602                       |
| Robotics (cs.RO)                 | 5,379                      | 7,933                       |
| Any of the above                 | 85,670                     | 107,380                     |

differences separate the two corpora. WoS includes some conference proceedings, but most of its papers (about 82%) are from journals. Its disciplinary coverage is also far broader, spanning fields in which we expect to find no AI-relevant papers. WoS includes 26.7M publications from 2010 through 2019, about 2.2M to 3.1M in each year.11

4 Learning from arXiv

Our baseline solution uses keyword matches to identify AI-relevant publications. We use 100 terms and patterns that we developed for a variety of document retrieval tasks in early Spring 2019, in a manual process: we reviewed search results and adapted the term list, and iterated until satisfied. (See Appendix A.) If one of these terms is present in the title or abstract of a publication, we consider that publication AI-relevant. Our expectation was that this approach would achieve reasonable precision but low recall. When tested against arXiv papers, considering papers in any of the six chosen subjects to be AI-relevant, we observe precision of .76 and recall of .43 ($F_1 = .55$).

A second approach for comparison is a keyword-classifier hybrid developed by Elsevier [22] as part of a bibliometric study of AI. The Elsevier group first extracted candidate terms from diverse textual sources, drawing from syllabi, books, patents, textbooks, the Cooperative Patent Classification scheme,12 and AI news coverage.13 The initial result was 800,000 keywords, which the group iteratively reduced to 797 distinct and specific terms.

The Elsevier team solicited comments on this set of terms from outside subject-matter experts. Characteristically,14 however, these experts could not agree on any common set of keywords “representative enough to scope the breadth of the field and […] specific enough to AI” [22]. The solution was for internal experts to score the terms on a three-point scale, and then task the outside experts with labeling a collection of publications that included the keywords. This account illustrates the difficulty of delineating the field by consensus, and the investment that expert labeling entails.

Ultimately, incidence of the 797 terms in the input text was the basis for a series of features: variously weighted counts and proportions of lower- and higher-scoring terms in title and abstract text. Following [22], we apply a random forest model to learn weights for these features using a training set drawn from the arXiv corpus.14

We depart from a replication of the Elsevier method by training on arXiv, and the implementation details of doing so may not correspond with the original work. Using a grid search to tune hyperparameter values and evaluating performance through cross-validation, we see precision of .74 and recall of .49 ($F_1 = .59$) in prediction of AI-relevant articles. These results outperform our baseline keyword solution.15

Lastly, we apply SciBERT [4], a BERT [7] model pre-trained on full text from Semantic Scholar then frozen and used to embed the title and abstract text of publications for classification. We first consider papers tagged with any of the

---

11 These counts describe the collection as queried using bulk data retrieved from Clarivate on January 28, 2020. Throughout, we restrict the data to records in the WoS Core Collection, with WoS-prefixed identifiers, and publication years from 2010 through 2019.
12 https://www.cooperativepatentclassification.org/cpcSchemeAndDefinitions
13 https://aitopics.org
14 We partition the arXiv corpus in an 80% training, 10% evaluation, and 10% test split, stratifying by publication year and class.
15 For implementation details and replication code, see https://github.com/georgetown-cset/ai-relevant-papers
six subjects to be AI-relevant and train a binary “all subjects” classifier. In evaluation on the arXiv test set, we find improvements from SciBERT over the previous methods, with precision of 0.83 and recall of 0.85 ($F_1 = 0.84$).

We also train classifiers for AI-relevant subjects separately, one-versus-all\(^\text{16}\). This effort is successful for the three subjects that correspond with well-defined application fields: NLP ($F_1 = 0.86$), Computer Vision ($F_1 = 0.84$) and Robotics ($F_1 = 0.75$).

We see lower performance from a Machine Learning subject model ($F_1 = 0.59$), which suffers from comparatively low recall. Learning the Multiagent Systems and Artificial Intelligence subjects is more challenging, and we leave it to future work. There are very few examples of Multiagent Systems papers (only 2,602, of which 985 have cs.MA as their primary subject). The Artificial Intelligence subject is larger (19,964 papers), but defined so as to exclude papers that belong in any of the other AI-relevant subjects. As described above, it “includes Expert Systems, Theorem Proving [...], Knowledge Representation, Planning, and Uncertainty in AI.” The result is that more than half the papers in the Artificial Intelligence subject are cross-posts from another subject, and we speculate that this leads to greater heterogeneity.

In Table 2 we summarize the performance of the baseline keyword solution, the Elsevier method, and the SciBERT models. The all-subjects SciBERT model outperforms the alternative methods in the test data, and in comparison with the keyword-reliant solutions, we find appealing the availability of real-time updates from new arXiv content and the straightforward decomposability of AI-relevant research into subjects like computer vision.

| Method                              | Precision | Recall | $F_1$ |
|-------------------------------------|-----------|--------|-------|
| CSET Keywords                       | .76       | .43    | .55   |
| Elsevier Keyword-classifier Hybrid\[22\] | .74       | .49    | .59   |
| SciBERT All Subjects                | \textbf{.83} | \textbf{.85} | \textbf{.84} |
| SciBERT Natural Language Processing (cs.CL) | .86       | .86    | .86   |
| Computer Vision (cs.CV)             | .87       | .81    | .84   |
| Machine Learning (cs.LG, stat.ML)   | .79       | .47    | .59   |
| Robotics (cs.RO)                    | .78       | .73    | .75   |

The keyword solution performs best in the year we developed it, 2019, with $F_1$ of 0.61, and declines steadily in prior years. (See Figure\[1\].) This variation is unsurprising in a fast-moving field. Elsevier’s model and the SciBERT all-subjects model exhibit the same pattern, but for different reasons. Higher performance from the supervised methods in more recent years is due in large part to longitudinal imbalance in the training data\[17\].

The appropriate response is context-sensitive, because the expansion of arXiv since 2010 is attributable to its popularity relative to traditional journals, the growth of the particular fields arXiv covers, and secular trends in research output. When training a classifier on arXiv for inference in WoS or elsewhere, one might seek the highest performance overall or prefer stable performance within strata meaningful in downstream analysis. We suggest comparing the performance of a single model to that of period-specific models if inference focuses on time-series measures.

\[\text{16}\] The all-subject AI-relevance model is trained and evaluated on the same data as the Elsevier solution. For each of the subject models, we use same split proportions but stratify by year and the corresponding binary class (e.g., positive if Computer Vision, else negative). Throughout, we use the same tuning parameters as reported for the text classification task in [4].

\[\text{17}\] It is also possible that classification in earlier years is more difficult than in recent years, or for that matter easier, but the imbalance confounds direct evaluation.
Figure 1: Higher performance from the supervised methods in more recent years is due in large part to longitudinal imbalance in the training data. Resampling or other strategies for imbalanced data can address this as appropriate for downstream analyses. The variation in keyword performance, by contrast, is the sign of a fast-moving field.

5 Generalization

Because we lack gold labels for straightforward estimation of the models’ performance in WoS, we evaluate their predictions by comparison to other indicators of subject information. Figure 2 shows the proportion of articles predicted relevant by each subject model in a set of AI and ML journals available in WoS.

The Artificial Intelligence subject model shows high rates of positive predictions across many of the journals: 96% of the articles in Journal of AI Research, for example, and 87% in Journal of Machine Learning Research. The ML model also identifies a large number of articles as relevant. By comparison, the positive predictions of the NLP, CV, and Robotics models are far more restricted to domain-specific journals.

Figure 3 shows the positive-prediction rates for conference proceedings. Here, the Multiagent Systems model identifies 93-94% of Autonomous Agents and Multiagent Systems articles as relevant, while the NLP model flags 98-99% of proceedings from the three NLP conferences. The AI model is once again less subject-specific, as is the ML model to a lesser degree.

Lastly, we report the proportion of each WoS category predicted relevant by the subject models. Figure 4 contains a set of fields represented to some degree on arXiv, primarily in quantitative disciplines. Although without labeled data we cannot report accuracy scores, the highest proportions are in plausible categories: robotics, imaging, remote sensing, computer science, and automation and control systems. We exclude from inference the WoS categories whose contents are unlikely to overlap with arXiv’s. (See Appendix B.)
Figure 2: SciBERT models predict high proportions of publications in AI and ML journals in WoS to be subject-relevant.

| Source of Journals                                      | AI     | NLP    | CV     | ML     | MA     | RO     |
|---------------------------------------------------------|--------|--------|--------|--------|--------|--------|
| Journal of AI Research                                  | 0.96   | 0.18   | 0.06   | 0.48   | 0.55   | 0.27   |
| Computational Linguistics                               | 0.97   | 0.98   | 0.04   | 0.69   | 0.14   | 0.01   |
| Computational Linguistics and Related Fields            | 0.92   | 0.89   | 0.08   | 0.58   | 0.03   | 0.03   |
| Computational Linguistics: Applications                 | 0.93   | 1.00   | 0.00   | 0.20   | 0.27   | 0.07   |
| Computer Vision and Image Understanding                 | 0.47   | 0.03   | 0.99   | 0.49   | 0.08   | 0.42   |
| Int'l Journal of Computer Vision                         | 0.58   | 0.04   | 0.99   | 0.53   | 0.04   | 0.41   |
| Pattern Recognition                                      | 0.58   | 0.06   | 0.96   | 0.73   | 0.06   | 0.13   |
| Journal of Machine Learning Research                    | 0.87   | 0.08   | 0.29   | 0.98   | 0.11   | 0.05   |
| IEEE Transactions on Neural Networks                     | 0.81   | 0.05   | 0.40   | 0.86   | 0.14   | 0.12   |
| IEEE Transactions on Neural Networks and Learning Systems| 0.77   | 0.04   | 0.42   | 0.83   | 0.18   | 0.17   |
| Neural Networks                                          | 0.62   | 0.06   | 0.42   | 0.72   | 0.18   | 0.14   |
| Robotics and Autonomous Systems                          | 0.59   | 0.02   | 0.48   | 0.29   | 0.63   | 0.98   |
| IEEE Transactions on Robotics                            | 0.35   | 0.01   | 0.28   | 0.12   | 0.46   | 0.99   |
| Advanced Robotics                                        | 0.45   | 0.03   | 0.29   | 0.14   | 0.41   | 0.98   |

Figure 3: In WoS conference proceedings, subject models predict varying proportions of publications to be subject-relevant.

| Source of Conference                                      | AI     | CL    | CV    | LG    | MA    | RO    |
|-----------------------------------------------------------|--------|-------|-------|-------|-------|-------|
| Chinese Computational Linguistics                         | 0.96   | 0.99  | 0.14  | 0.71  | 0.85  | 0.00  |
| Computational Linguistics and Intelligent Text Processing| 0.98   | 0.99  | 0.04  | 0.57  | 0.09  | 0.01  |
| Association for Computational Linguistics                 | 0.98   | 0.98  | 0.03  | 0.70  | 0.11  | 0.02  |
| Int'l Joint Conf. on Neural Networks                      | 0.91   | 0.12  | 0.44  | 0.85  | 0.20  | 0.16  |
| Int'l Joint Conf. on AI                                   | 0.92   | 0.22  | 0.19  | 0.55  | 0.45  | 0.20  |
| Symposium on Applied Machine Intelligence and Informatics| 0.56   | 0.11  | 0.17  | 0.18  | 0.42  | 0.34  |
| Int'l Symposium on Applied Machine Intelligence and Informatics| 0.55 | 0.09  | 0.20  | 0.18  | 0.34  | 0.34  |
| Int'l Conf. on Autonomous Agents and Multiagent Systems   | 0.93   | 0.13  | 0.08  | 0.53  | 0.83  | 0.44  |
| Autonomous Agents and Multiagent Systems                  | 0.92   | 0.14  | 0.10  | 0.43  | 0.94  | 0.44  |
Figure 4: Where WoS and arXiv share topical coverage, we observe plausible rates of predicted subject relevance. Each row in the table represents publications in a WoS category, and each column a subject model. Cells give the proportion of publications in a category predicted relevant by a model. From left to right, the arXiv subject abbreviations refer to Artificial Intelligence, Computation and Language (NLP), Computer Vision, Machine Learning, Multiagent Systems, and Robotics.

| Category                             | AI  | CL  | CV  | LG  | MA  | RO  |
|--------------------------------------|-----|-----|-----|-----|-----|-----|
| Acoustics                            | 0.14| 0.13| 0.15| 0.19| 0.04| 0.08|
| Astronomy & Astrophysics             | 0.00| 0.00| 0.01| 0.01| 0.00| 0.01|
| Automation & Control Systems         | 0.31| 0.02| 0.12| 0.15| 0.32| 0.44|
| Biochemistry & Molecular Biology     | 0.02| 0.01| 0.02| 0.03| 0.01| 0.00|
| Biophysics                           | 0.03| 0.00| 0.10| 0.04| 0.02| 0.10|
| Biotechnology & Applied Microbiology | 0.02| 0.01| 0.01| 0.02| 0.01| 0.01|
| Cell Biology                         | 0.01| 0.01| 0.02| 0.01| 0.00| 0.00|
| Chemistry                            | 0.01| 0.00| 0.01| 0.01| 0.01| 0.00|
| Computer Science                     | 0.45| 0.10| 0.21| 0.24| 0.28| 0.13|
| Developmental Biology                | 0.01| 0.01| 0.04| 0.01| 0.01| 0.01|
| Energy & Fuels                       | 0.07| 0.00| 0.01| 0.04| 0.11| 0.04|
| Engineering                          | 0.10| 0.01| 0.08| 0.06| 0.09| 0.11|
| Genetics & Heredity                  | 0.02| 0.02| 0.01| 0.03| 0.01| 0.00|
| Geochemistry & Geophysics            | 0.02| 0.00| 0.07| 0.03| 0.01| 0.01|
| Geology                              | 0.02| 0.00| 0.06| 0.02| 0.04| 0.01|
| Imaging Science & Photographic Technology | 0.26| 0.02| 0.66| 0.18| 0.05| 0.12|
| Information Science & Library Science | 0.30| 0.26| 0.03| 0.04| 0.39| 0.01|
| Instruments & Instrumentation        | 0.04| 0.00| 0.08| 0.04| 0.02| 0.05|
| Life Sciences & Biomedicine - Other Topics | 0.06| 0.02| 0.05| 0.05| 0.10| 0.03|
| Marine & Freshwater Biology          | 0.00| 0.00| 0.02| 0.01| 0.05| 0.01|
| Materials Science                    | 0.01| 0.00| 0.01| 0.01| 0.01| 0.02|
| Mathematical & Computational Biology | 0.33| 0.06| 0.18| 0.30| 0.11| 0.06|
| Mathematics                          | 0.06| 0.01| 0.02| 0.06| 0.04| 0.01|
| Mechanics                            | 0.01| 0.00| 0.01| 0.01| 0.03| 0.06|
| Metallurgy & Metallurgical Engineering | 0.02| 0.00| 0.01| 0.01| 0.03| 0.03|
| Meteorology & Atmospheric Sciences   | 0.01| 0.00| 0.02| 0.03| 0.02| 0.00|
| Microbiology                         | 0.00| 0.01| 0.00| 0.00| 0.00| 0.00|
| Microscopy                           | 0.01| 0.00| 0.17| 0.02| 0.00| 0.01|
| Mineralogy                           | 0.01| 0.00| 0.01| 0.00| 0.01| 0.00|
| Neurosciences & Neurology            | 0.11| 0.04| 0.15| 0.04| 0.02| 0.07|
| Nuclear Science & Technology         | 0.03| 0.00| 0.02| 0.01| 0.02| 0.01|
| Optics                               | 0.02| 0.00| 0.12| 0.02| 0.01| 0.02|
| Physics                              | 0.01| 0.00| 0.01| 0.01| 0.01| 0.01|
| Remote Sensing                       | 0.13| 0.00| 0.55| 0.17| 0.06| 0.12|
| Robotics                             | 0.47| 0.03| 0.31| 0.16| 0.46| 0.91|
| Thermodynamics                       | 0.02| 0.00| 0.01| 0.01| 0.03| 0.02|
6 Conclusion

Our results demonstrate high classification performance from SciBERT [4] models applied to learning arXiv subjects. Although we did not evaluate SciBERT against a comparable BERT model pre-trained on Wikipedia and the BookCorpus [7], we attribute some of this performance to transfer learning via SciBERT’s embedding of scientific vocabulary after pre-training on Semantic Scholar. Within the set of topics the models saw in training on arXiv papers, inference in WoS appears feasible: we observe plausible rates of predicted relevance in conference proceedings, journal articles, and the journal-based categories in WoS.

Looking forward, manual annotation is the obvious solution to our lack of labeled examples in WoS. However, developing guidelines for labeling publications for AI-relevance would require addressing definitional questions we sidestepped in this work; it would represent a departure from using the implicit delineation of the field provided by arXiv preprints. But we anticipate that labeling examples to approximate the boundaries of arXiv subjects, like NLP and computer vision, is far more tractable than manual labeling for AI relevance.

The arXiv corpus exhibits a class imbalance of about 9:1 in favor of negative examples. In WoS, whose topical coverage is broader, we assume the true imbalance is greater. The appropriate tuning for class performance will depend on the application.

Another major direction for future work is expanding domain generalizibility, particularly in potential application areas. We have substantive interest in papers on topics unavailable in arXiv, from agriculture to medicine. We would consider reports of AI applications in trade journals to be AI-relevant in principle, for example, but we focus in this paper on a delineation of the field whose implementation may not include them. To expand into these areas, we anticipate leveraging bibliometric data in addition to text: applying scientometric methods to extend the identification of publications describing the development and applications of AI beyond arXiv’s coverage.

References

[1] Artificial intelligence: How knowledge is created, transferred, and used. URL https://www.elsevier.com/research-intelligence/resource-library/ai-report

[2] G. C. Allen. Understanding china’s AI strategy: Clues to chinese strategic thinking on artificial intelligence and national security.

[3] S. K. Arora, A. L. Porter, J. Youtie, and P. Shapira. Capturing new developments in an emerging technology: An updated search strategy for identifying nanotechnology research outputs. Scientometrics, 95(1):351–370, Apr. 2013. ISSN 1588-2861. doi: 10.1007/s11192-012-0903-6.

[4] I. Beltagy, K. Lo, and A. Cohan. Scibert: Pretrained language model for scientific text. In EMNLP, 2019.

[5] K. W. Boyack, M. Patek, L. H. Ungar, P. Yoon, and R. Klavans. Classification of individual articles from all of science by research level. 8(1):1–12. ISSN 1751-1577. doi: 10.1016/j.joi.2013.10.005. URL http://www.sciencedirect.com/science/article/pii/S1751157713000825

[6] C. B. Clement, M. Bierbaum, K. P. O’Keeffe, and A. A. Alemi. On the use of ArXiv as a dataset. URL http://arxiv.org/abs/1905.00075 version: 1.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

[8] F. Gao, X. Jia, Z. Zhao, C.-C. Chen, F. Xu, Z. Geng, and X. Song. Bibliometric analysis on tendency and topics of artificial intelligence over last decade. ISSN 0946-7076, 1432-1858. doi: 10.1007/s00542-019-04426-y. URL http://link.springer.com/10.1007/s00542-019-04426-y

[9] C. Huang, A. Notten, and N. Rasters. Nanoscience and technology publications and patents: A review of social science studies and search strategies. The Journal of Technology Transfer, 36(2):145–172, Apr. 2011. ISSN 1573-7047. doi: 10.1007/s10961-009-9149-8.

[10] Y. Huang, J. Schuehle, A. L. Porter, and J. Youtie. A systematic method to create search strategies for emerging technologies based on the web of science; illustrated for ‘big data’. 105(3):2005–2022. ISSN 0138-9130, 1588-2861. doi: 10.1007/s11192-015-1638-y. URL http://link.springer.com/10.1007/s11192-015-1638-y
[11] Y. Huang, J. Schuehle, A. L. Porter, and J. Youtie. A systematic method to create search strategies for emerging technologies based on the Web of Science: Illustrated for ‘Big Data’. *Scientometrics*, 105(3):2005–2022, Dec. 2015. ISSN 0138-9130, 1588-2861. doi: 10.1007/s11192-015-1638-y.

[12] P. M. Krafft, M. Young, M. Katell, K. Huang, and G. Bugingo. Defining AI in policy versus practice. URL [http://arxiv.org/abs/1912.11095](http://arxiv.org/abs/1912.11095).

[13] F. Martínez-Plumed, B. S. Loe, P. Flach, S. Ó hÉigeartaigh, K. Vold, and J. Hernández-Orallo. The Facets of Artificial Intelligence: A Framework to Track the Evolution of AI. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pages 5180–5187, Stockholm, Sweden, July 2018. International Joint Conferences on Artificial Intelligence Organization. ISBN 978-0-992411-2-7. doi: 10.24963/ijcai.2018/718.

[14] D. H. Milanez, E. Noyons, and L. I. L. de Faria. A delineating procedure to retrieve relevant publication data in research areas: the case of nanocellulose. 107(2):627–643. ISSN 1588-2861. doi: 10.1007/s11192-016-1922-5. URL [https://doi.org/10.1007/s11192-016-1922-5](https://doi.org/10.1007/s11192-016-1922-5).

[15] K. Miyazaki and R. Sato. Analyses of the technological accumulation over the 2 nd and the 3 rd AI boom and the issues related to AI adoption by firms. In *2018 Portland International Conference on Management of Engineering and Technology (PICMET)*, pages 1–7. IEEE. ISBN 978-1-890843-37-3. doi: 10.23919/PICMET.2018.8481822. URL [https://ieeexplore.ieee.org/document/8481822/](https://ieeexplore.ieee.org/document/8481822/).

[16] A. Mogoutov and B. Kahane. Data search strategy for science and technology emergence: A scalable and evolutionary query for nanotechnology tracking. 36(6):893–903. ISSN 00487333. doi: 10.1016/j.respol.2007.02.005. URL [https://linkinghub.elsevier.com/retrieve/pii/S0048733307000479](https://linkinghub.elsevier.com/retrieve/pii/S0048733307000479).

[17] J. Niu, W. Tang, F. Xu, X. Zhou, and Y. Song. Global research on artificial intelligence from 1990–2014: Spatially-explicit bibliometric analysis. 5(5):66. ISSN 2220-9964. doi: 10.3390/ijgi5050066. URL [http://www.mdpi.com/2220-9964/5/5/66](http://www.mdpi.com/2220-9964/5/5/66).

[18] Raymond Perrault, Yoav Shoham, Erik Brynjolfsson, Jack Clark, John Etchemendy, Barbara Grosz, Terah Lyons, James Manyika, Saurabh Mishra, and Juan Carlos Niebles. The AI index 2019 annual report. URL [https://hai.stanford.edu/sites/g/files/sbiybj10986/f/ai_index_2019_report.pdf](https://hai.stanford.edu/sites/g/files/sbiybj10986/f/ai_index_2019_report.pdf).

[19] J. Rincon-Patino, G. Ramirez-Gonzalez, and J. C. Corrales. Exploring machine learning: A bibliometric general approach using Citespace. *F1000Research*, 7:1240, Aug. 2018. ISSN 2046-1402. doi: 10.12688/f1000research.15619.1.

[20] S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Pearson, 3 edition edition. ISBN 978-0-13-604259-4.

[21] A. K. Shukla, M. Janmaijaya, A. Abraham, and P. K. Muhuri. Engineering applications of artificial intelligence: A bibliometric analysis of 30 years (1988–2018). *Engineering Applications of Artificial Intelligence*, 85:517–532, Oct. 2019. ISSN 09521976. doi: 10.1016/j.engappai.2019.06.010.

[22] M. Siebert, C. Kohler, A. Scerri, and G. Tsatsaronis. Technical background and methodology for the elsevier’s artificial intelligence report.

[23] H. Small, K. W. Boyack, and R. Klavans. Identifying emerging topics in science and technology. 43(8):1450–1467. ISSN 00487333. doi: 10.1016/j.respol.2014.02.005. URL [https://linkinghub.elsevier.com/retrieve/pii/S0048733314000298](https://linkinghub.elsevier.com/retrieve/pii/S0048733314000298).

[24] A. Suominen and N. C. Newman. Exploring the Fundamental Conceptual Units of Technical Emergence. In *2017 Portland International Conference on Management of Engineering and Technology (PICMET)*, pages 1–5, July 2017. doi: 10.23919/PICMET.2017.8125287.

[25] D. Tarraf, W. Shelton, E. Parker, B. Alkire, D. Carey, J. Grana, A. Levedahl, J. Leveille, J. Mondschein, J. Ryseff, A. Wyne, D. Elinoff, E. Geist, B. Harris, E. Hui, C. Kenney, S. Newberry, C. Sachs, P. Schirmer, D. Schlang, V. Smith, A. Tingstad, P. Vedula, and K. Warren. The Department of Defense Posture for Artificial Intelligence: Assessment and Recommendations. RAND Corporation. ISBN 978-1-977404-05-3. doi: 10.7249/RR4229. URL [https://www.rand.org/pubs/research_reports/RR4229.html](https://www.rand.org/pubs/research_reports/RR4229.html).

[26] M. Zitt. Meso-level retrieval: IR-bibliometrics interplay and hybrid citation-words methods in scientific fields delineation. 102(3):2223–2245. ISSN 1588-2861. doi: 10.1007/s11192-014-1482-5. URL [https://doi.org/10.1007/s11192-014-1482-5](https://doi.org/10.1007/s11192-014-1482-5).
### Keywords

Table A.1: We use these terms and patterns in our baseline search strategy. Originally, we developed this list for document retrieval tasks on a variety of knowledgebases, such as WoS, ProQuest, Dimensions, and CNKI, in early Spring 2019. The * character represents a wildcard that matches zero or more non-whitespace characters.

| Active Learning | Incremental Clustering | Adaptive Learning | Information Extraction | Anomaly Detection | Information Fusion | Artificial Intelligence | Information Retrieval | Associative Learning | K Nearest Neighbor |
|----------------|------------------------|------------------|------------------------|------------------|-------------------|------------------------|----------------------|---------------------|-------------------|
| Autonomous Navigation | Knowledge Based System* | Autonomous System* | Knowledge Discovery | Autonomous Vehicle* | Knowledge Representation | Average Link Clustering | Language Identification | Back Propagation | Machine Learning |
| Backpropagation | Machine Perception | Binary Classification | Machine Translation | BioNLP | Multi Class Classification | Character Recognition | Multi Task Learning | Classification Algorithm | Natural Language Generation |
| Classification Algorithm | Natural Language Processing | Classification Label* | Natural Language Understanding | Clustering Method* | Neural Network | Complete Link Clustering | Object Recognition | Computer Aided Diagnosis | One Shot Learning |
| Computer Vision | Pattern Matching | Deep Learning | Pattern Recognition | Ensemble Learning | Random Forest | Evolutionary Algorithm | Recommendation System* | Fac Expression Recognition | Recurrent Network |
| Fac Identification | Reinforcement Learning | Fac Recognition | Scene* Classification | Feature Extraction | Scene* Understanding | Feature Learning | Self Driving Car* | Feature Matching | Semi Supervised Learning |
| Feature Selection | Sentiment Classification | Feature Vector | Image Classification | Feedback Forward Network | Spatial Learning | Fuzzy Clustering | Speech Processing | Gradient Algorithm | Speech Recognition |
| Generative Adversarial Network | Speech Synthesis | Graph Matching | Statistical Learning | Graphical Model | Strong Artificial Intelligence | Handwriting Recognition | Supervised Learning | Hierarchical Clustering | Support Vector Machine |
| Hierarchical Model | Text Mining | Human Robot | Text Processing | Image Annotation | Transfer Learning | Image Classification | Translation System | Image Matching | Unsupervised Learning |
| Image Classification | Video Classification | Image Matching | Video Processing | Image Processing | Image Registration | Image Representation | Weak Artificial Intelligence | Image Retrieval | Zero Shot Learning |
## arXiv coverage

Table B.1: We limit inference to the WoS categories listed below whose contents are likely to overlap with arXiv’s. Category names are given verbatim, as they appear in Clarivate data. Typology values refer to WoS `ascatype`.

| Category                                      | Typology  | Count  |
|-----------------------------------------------|-----------|--------|
| Acoustics                                     | Traditional | 154,404 |
| Acoustics                                     | Extended   | 154,703 |
| Astronomy & Astrophysics                       | Traditional | 467,534 |
| Astronomy & Astrophysics                       | Extended   | 479,191 |
| Astronomy Astrophysics                         | Traditional | 7,763  |
| Automation & Control Systems                   | Traditional | 492,410 |
| Automation & Control Systems                   | Extended   | 627,700 |
| Automation Control Systems                     | Traditional | 129,647 |
| Biochemistry & Molecular Biology               | Traditional | 1,450,357 |
| Biochemistry & Molecular Biology               | Extended   | 1,731,807 |
| Biology                                        | Traditional | 388,531 |
| Biophysics                                     | Traditional | 375,695 |
| Biophysics                                     | Extended   | 378,238 |
| Biotechnology & Applied Microbiology           | Traditional | 571,706 |
| Biotechnology & Applied Microbiology           | Extended   | 586,875 |
| Biotechnology Applied Microbiology             | Traditional | 9,965  |
| Business & Economics                           | Extended   | 1,279,720 |
| Cell Biology                                   | Traditional | 829,144 |
| Cell Biology                                   | Extended   | 858,881 |
| Computer Science                               | Extended   | 2,955,517 |
| Computer Science Cybernetics                   | Traditional | 3,001  |
| Computer Science Interdisciplinary Applications| Traditional | 295,152 |
| Computer Science, Artificial Intelligence      | Traditional | 868,308 |
| Computer Science, Cybernetics                  | Traditional | 100,954 |
| Computer Science, Hardware & Architecture      | Traditional | 322,719 |
| Computer Science, Information Systems          | Traditional | 759,069 |
| Computer Science, Interdisciplinary Applications| Traditional | 534,590 |
| Computer Science, Software Engineering         | Traditional | 392,885 |
| Computer Science, Theory & Methods             | Traditional | 897,664 |
| Construction Building Technology               | Traditional | 112,822 |
| Economics                                      | Traditional | 651,087 |
| Energy Fuels                                   | Traditional | 108,762 |
| Engineering Aerospace                          | Traditional | 60,964  |
| Engineering Biomedical                         | Traditional | 3,472  |
| Engineering Chemical                           | Traditional | 166,717 |
| Engineering Electrical Electronic              | Traditional | 317,137 |
| Engineering Industrial                         | Traditional | 83,927  |
| Engineering Manufacturing                      | Traditional | 20,239  |
| Engineering Mechanical                         | Traditional | 76,829  |
| Engineering Multidisciplinary                   | Traditional | 49,358  |
| Engineering, Aerospace                         | Traditional | 138,674 |
| Engineering, Biomedical                        | Traditional | 343,099 |
| Engineering, Chemical                          | Traditional | 615,237 |
| Engineering, Civil                             | Traditional | 388,419 |
| Engineering, Electrical & Electronic           | Traditional | 2,667,329 |
| Engineering, Environmental                     | Traditional | 295,582 |
| Engineering, Geological                        | Traditional | 95,575  |
| Engineering, Industrial                        | Traditional | 204,533 |
| Engineering, Manufacturing                     | Traditional | 228,759 |
| Engineering, Marine                            | Traditional | 40,723  |
| Engineering, Mechanical                        | Traditional | 647,335 |
Table B.1: We limit inference to the WoS categories listed below whose contents are likely to overlap with arXiv’s (continued).

| Category                                                      | Typology  | Count    |
|---------------------------------------------------------------|-----------|----------|
| Engineering, Multidisciplinary                                | Traditional | 344,869  |
| Engineering, Ocean                                             | Traditional | 46,010   |
| Engineering, Petroleum                                         | Traditional | 68,667   |
| Genetics Heredity                                              | Traditional | 37,373   |
| Geochemistry & Geophysics                                      | Traditional | 223,783  |
| Geochemistry & Geophysics                                      | Extended  | 232,184  |
| Green & Sustainable Science & Technology                       | Traditional | 146,852  |
| Imaging Science & Photographic Technology                      | Traditional | 251,296  |
| Imaging Science & Photographic Technology                      | Extended  | 253,317  |
| Imaging Science Photographic Technology                        | Traditional | 836      |
| Information Science & Library Science                          | Traditional | 251,050  |
| Information Science & Library Science                          | Extended  | 263,526  |
| Information Science Library Science                            | Traditional | 7,911    |
| Instruments & Instrumentation                                  | Traditional | 408,552  |
| Instruments & Instrumentation                                  | Extended  | 415,074  |
| Language & Linguistics                                         | Traditional | 188,787  |
| Language Linguistics                                           | Traditional | 794      |
| Life Sciences & Biomedicine - Other Topics                     | Extended  | 391,636  |
| Logic                                                          | Traditional | 23,145   |
| Materials Science                                              | Extended  | 2,415,451|
| Materials Science Multidisciplinary                            | Traditional | 132      |
| Materials Science Paper Wood                                   | Traditional | 2,117    |
| Materials Science Textiles                                     | Traditional | 1,607    |
| Materials Science, Biomaterials                                | Traditional | 113,911  |
| Materials Science, Ceramics                                    | Traditional | 138,510  |
| Materials Science, Characterization & Testing                 | Traditional | 106,330  |
| Materials Science, Coatings & Films                           | Traditional | 157,952  |
| Materials Science, Composites                                  | Traditional | 100,521  |
| Materials Science, Multidisciplinary                           | Traditional | 1,817,377|
| Materials Science, Paper & Wood                               | Traditional | 45,318   |
| Materials Science, Textiles                                    | Traditional | 52,610   |
| Mathematical & Computational Biology                           | Traditional | 142,131  |
| Mathematical & Computational Biology                           | Extended  | 153,402  |
| Mathematical Computational Biology                             | Traditional | 10,940   |
| Mathematical Methods In Social Sciences                        | Extended  | 61,831   |
| Mathematics                                                    | Traditional | 589,478  |
| Mathematics                                                    | Extended  | 1,329,236|
| Mathematics Applied                                            | Traditional | 22,906   |
| Mathematics, Applied                                           | Traditional | 527,864  |
| Mathematics, Interdisciplinary Applications                    | Traditional | 178,185  |
| Mechanics                                                      | Traditional | 446,852  |
| Mechanics                                                      | Extended  | 460,610  |
| Meteorology & Atmospheric Sciences                             | Traditional | 234,653  |
| Meteorology & Atmospheric Sciences                             | Extended  | 272,158  |
| Meteorology Atmospheric Sciences                               | Traditional | 34,535   |
| Microbiology                                                    | Traditional | 457,169  |
| Microbiology                                                    | Extended  | 460,385  |
| Microscopy                                                      | Traditional | 33,121   |
| Microscopy                                                      | Extended  | 33,137   |
| Nanoscience & Nanotechnology                                  | Traditional | 547,640  |
| Neuroimaging                                                    | Traditional | 52,795   |
| Neurosciences                                                   | Traditional | 1,066,397|
| Neurosciences & Neurology                                      | Extended  | 1,756,544|
| Nuclear Science & Technology                                   | Traditional | 215,737  |
| Nuclear Science & Technology                                   | Extended  | 240,096  |
Table B.1: We limit inference to the WoS categories listed below whose contents are likely to overlap with arXiv’s (continued).

| Category                          | Typology     | Count       |
|-----------------------------------|--------------|-------------|
| Nuclear Science Technology        | Traditional  | 21,181      |
| Optics                            | Traditional  | 870,400     |
| Optics                            | Extended     | 879,145     |
| Physics                           | Extended     | 3,189,791   |
| Physics Applied                   | Traditional  | 457         |
| Physics Atomic Molecular Chemical | Traditional  | 4,449       |
| Physics Multidisciplinary          | Traditional  | 109,509     |
| Physics Nuclear                   | Traditional  | 6,070       |
| Physics, Applied                  | Traditional  | 1,388,092   |
| Physics, Atomic, Molecular & Chemical | Traditional | 327,528   |
| Physics, Condensed Matter         | Traditional  | 665,465     |
| Physics, Fluids & Plasmas         | Traditional  | 182,213     |
| Physics, Mathematical             | Traditional  | 215,522     |
| Physics, Multidisciplinary        | Traditional  | 505,808     |
| Physics, Nuclear                  | Traditional  | 181,227     |
| Physics, Particles & Fields       | Traditional  | 278,522     |
| Plant Sciences                    | Traditional  | 561,836     |
| Plant Sciences                    | Extended     | 568,219     |
| Polymer Science                   | Traditional  | 358,720     |
| Polymer Science                   | Extended     | 360,136     |
| Quantum Science & Technology      | Traditional  | 40,025      |
| Remote Sensing                    | Traditional  | 197,070     |
| Remote Sensing                    | Extended     | 200,094     |
| Robotics                          | Traditional  | 168,571     |
| Robotics                          | Extended     | 174,071     |
| Science & Technology - Other Topics | Extended   | 1,616,680   |
| Spectroscopy                      | Traditional  | 149,683     |
| Spectroscopy                      | Extended     | 151,151     |
| Statistics & Probability          | Traditional  | 199,849     |
| Telecommunications                | Traditional  | 776,565     |
| Telecommunications                | Extended     | 779,352     |
| Thermodynamics                    | Traditional  | 231,418     |
| Thermodynamics                    | Extended     | 234,360     |
| Zoology                           | Traditional  | 294,619     |
| Zoology                           | Extended     | 326,130     |

Table B.2: We exclude from inference the WoS categories listed below whose contents are not likely to overlap with arXiv’s. Category names are given verbatim, as they appear in Clarivate data. Typology refers to WoS ascatype.

| Category                         | Typology     | Count       |
|----------------------------------|--------------|-------------|
| Agricultural Economics & Policy  | Traditional  | 29,709      |
| Agricultural Economics Policy    | Traditional  | 26,124      |
| Agricultural Engineering         | Traditional  | 82,333      |
| Agriculture                      | Extended     | 1,121,220   |
| Agriculture Multidisciplinary    | Traditional  | 174,550     |
| Agriculture, Dairy & Animal Science | Traditional | 180,338   |
| Agriculture, Multidisciplinary   | Traditional  | 138,433     |
| Agronomy                         | Traditional  | 304,306     |
| Allergy                          | Traditional  | 112,097     |
| Allergy                          | Extended     | 113,863     |
| Anatomy & Morphology             | Traditional  | 43,775      |
| Anatomy & Morphology             | Extended     | 47,082      |
Table B.2: We exclude from inference the WoS categories listed below whose contents are not likely to overlap with arXiv’s (continued).

| Category                                      | Typology   | Count  |
|-----------------------------------------------|------------|--------|
| Anatomy Morphology                            | Traditional| 581    |
| Andrology                                     | Traditional| 14,405 |
| Anesthesiology                               | Traditional| 154,360|
| Anesthesiology                               | Extended   | 155,534|
| Anthropology                                 | Traditional| 148,928|
| Anthropology                                 | Extended   | 151,542|
| Archaeology                                  | Traditional| 86,286 |
| Archaeology                                  | Extended   | 91,422 |
| Architecture                                 | Traditional| 110,406|
| Architecture                                 | Extended   | 110,518|
| Area Studies                                  | Traditional| 144,905|
| Area Studies                                  | Extended   | 151,249|
| Art                                           | Traditional| 165,849|
| Art                                           | Extended   | 166,135|
| Arts & Humanities - Other Topics              | Extended   | 425,537|
| Asian Studies                                 | Traditional| 60,949 |
| Asian Studies                                 | Extended   | 63,839 |
| Audiology & Speech-Language Pathology         | Extended   | 47,136 |
| Audiology & Speech-Language Pathology         | Traditional| 47,136 |
| Behavioral Sciences                           | Traditional| 143,265|
| Behavioral Sciences                           | Extended   | 143,587|
| Biochemical Research Methods                  | Traditional| 332,781|
| Biochemistry Molecular Biology                | Traditional| 25,466 |
| Biodiversity & Conservation                   | Extended   | 100,000|
| Biodiversity Conservation                     | Traditional| 97,899 |
| Biomedical Social Sciences                    | Extended   | 85,753 |
| Business                                      | Traditional| 289,968|
| Business Finance                              | Traditional| 14,419 |
| Business, Finance                             | Traditional| 162,612|
| Cardiac & Cardiovascular Systems              | Traditional| 804,110|
| Cardiac Cardiovascular Systems                | Traditional| 574    |
| Cardiovascular System & Cardiology           | Extended   | 1,109,975|
| Cell & Tissue Engineering                     | Traditional| 55,696 |
| Chemistry                                     | Extended   | 3,712,333|
| Chemistry Analytical                          | Traditional| 59,982 |
| Chemistry Applied                             | Traditional| 581    |
| Chemistry Multidisciplinary                    | Traditional| 67,157 |
| Chemistry Organic                             | Traditional| 28,163 |
| Chemistry Physical                            | Traditional| 51,751 |
| Chemistry, Analytical                         | Traditional| 442,433|
| Chemistry, Applied                            | Traditional| 272,801|
| Chemistry, Inorganic & Nuclear                | Traditional| 255,875|
| Chemistry, Medicinal                          | Traditional| 261,868|
| Chemistry, Multidisciplinary                   | Traditional| 1,428,721|
| Chemistry, Organic                            | Traditional| 396,120|
| Chemistry, Physical                           | Traditional| 1,024,973|
| Classics                                      | Traditional| 61,230 |
| Classics                                      | Extended   | 62,474 |
| Clinical Neurology                            | Traditional| 1,008,455|
| Communication                                 | Traditional| 122,860|
| Communication                                 | Extended   | 124,943|
| Construction & Building Technology            | Traditional| 194,776 |
| Construction & Building Technology            | Extended   | 315,958 |
| Criminology & Penology                        | Traditional| 69,699 |
| Criminology & Penology                        | Extended   | 70,576 |

15
Table B.2: We exclude from inference the WoS categories listed below whose contents are not likely to overlap with arXiv’s (continued).

| Category                                      | Typology     | Count       |
|-----------------------------------------------|--------------|-------------|
| Critical Care Medicine                        | Traditional  | 253,975     |
| Crystallography                               | Traditional  | 197,470     |
| Crystallography                               | Extended     | 197,779     |
| Cultural Studies                              | Traditional  | 57,679      |
| Cultural Studies                              | Extended     | 59,852      |
| Dance                                         | Extended     | 36,111      |
| Dance                                         | Traditional  | 36,111      |
| Demography                                    | Traditional  | 34,511      |
| Demography                                    | Extended     | 36,672      |
| Dentistry Oral Surgery Medicine               | Traditional  | 22,677      |
| Dentistry, Oral Surgery & Medicine            | Traditional  | 224,355     |
| Dentistry, Oral Surgery & Medicine            | Extended     | 249,707     |
| Dermatology                                   | Traditional  | 286,332     |
| Dermatology                                   | Extended     | 288,882     |
| Developmental Studies                         | Extended     | 46,520      |
| Developmental Studies                         | Traditional  | 46,520      |
| Developmental Biology                         | Traditional  | 114,300     |
| Developmental Biology                         | Extended     | 116,108     |
| Ecology                                       | Traditional  | 371,343     |
| Education & Educational Research              | Traditional  | 529,168     |
| Education & Educational Research              | Extended     | 680,855     |
| Education, Scientific Disciplines            | Traditional  | 7,282       |
| Education, Special                           | Traditional  | 530         |
| Emergency Medicine                            | Traditional  | 90,582      |
| Emergency Medicine                            | Extended     | 93,874      |
| Endocrinology & Metabolism                    | Traditional  | 579,750     |
| Endocrinology & Metabolism                    | Extended     | 595,991     |
| Endocrinology Metabolism                      | Traditional  | 195         |
| Energy & Fuels                                | Traditional  | 632,499     |
| Energy & Fuels                                | Extended     | 749,388     |
| Engineering                                   | Extended     | 6,344,262   |
| Entomology                                    | Traditional  | 139,402     |
| Entomology                                    | Extended     | 142,530     |
| Environmental Sciences                        | Traditional  | 1,009,900   |
| Environmental Sciences & Ecology              | Extended     | 1,474,360   |
| Environmental Studies                         | Traditional  | 206,046     |
| Ergonomics                                    | Traditional  | 33,112      |
| Ethics                                        | Traditional  | 76,246      |
| Ethnic Studies                                | Traditional  | 43,473      |
| Ethnic Studies                                | Extended     | 48,784      |
| Evolutionary Biology                          | Extended     | 136,263     |
| Evolutionary Biology                          | Traditional  | 136,263     |
| Family Studies                                | Extended     | 58,526      |
| Family Studies                                | Traditional  | 58,526      |
| Film Radio Television                         | Traditional  | 55          |
| Film, Radio & Television                      | Extended     | 98,276      |
| Film, Radio, Television                       | Traditional  | 98,221      |
| Fisheries                                     | Traditional  | 137,544     |
| Fisheries                                     | Extended     | 137,655     |
| Folklore                                      | Traditional  | 17,117      |
| Food Science & Technology                     | Traditional  | 433,732     |
Table B.2: We exclude from inference the WoS categories listed below whose contents are not likely to overlap with arXiv’s (continued).

| Category                                      | Typology     | Count     |
|-----------------------------------------------|--------------|-----------|
| Food Science & Technology                     | Extended     | 516,086   |
| Food Science Technology                       | Traditional  | 71,329    |
| Forestry                                      | Traditional  | 151,593   |
| Forestry                                      | Extended     | 153,966   |
| Gastroenterology & Hepatology                 | Traditional  | 578,127   |
| Gastroenterology & Hepatology                 | Extended     | 578,877   |
| Gastroenterology Hepatology                   | Traditional  | 160       |
| General & Internal Medicine                   | Extended     | 1,981,284 |
| Genetics & Heredity                           | Traditional  | 497,124   |
| Genetics & Heredity                           | Extended     | 538,484   |
| Geography                                     | Traditional  | 119,322   |
| Geography                                     | Extended     | 125,556   |
| Geography, Physical                           | Traditional  | 108,154   |
| Geology                                       | Traditional  | 120,300   |
| Geology                                       | Extended     | 639,757   |
| Geosciences Multidisciplinary                 | Traditional  | 71,444    |
| Geosciences, Multidisciplinary                | Traditional  | 441,098   |
| Geriatrics & Gerontology                      | Traditional  | 144,277   |
| Geriatrics & Gerontology                      | Extended     | 201,882   |
| Gerontology                                   | Traditional  | 127,831   |
| Government & Law                              | Extended     | 612,045   |
| Health Care Sciences & Services               | Traditional  | 298,787   |
| Health Care Sciences & Services               | Extended     | 361,661   |
| Health Policy & Services                      | Traditional  | 184,094   |
| Hematology                                    | Traditional  | 511,342   |
| Hematology                                    | Extended     | 511,925   |
| History                                       | Traditional  | 602,685   |
| History                                       | Extended     | 616,148   |
| History & Philosophy Of Science               | Traditional  | 109,953   |
| History & Philosophy Of Science               | Extended     | 110,244   |
| History Of Social Sciences                    | Traditional  | 41,988    |
| Horticulture                                  | Traditional  | 164,497   |
| Hospitality, Leisure, Sport & Tourism         | Traditional  | 92,321    |
| Humanities Multidisciplinary                  | Traditional  | 518       |
| Humanities, Multidisciplinary                 | Traditional  | 310,361   |
| Immunology                                    | Traditional  | 686,286   |
| Immunology                                    | Extended     | 690,647   |
| Industrial Relations & Labor                  | Traditional  | 40,123    |
| Industrial Relations Labor                    | Traditional  | 297       |
| Infectious Diseases                           | Traditional  | 334,215   |
| Infectious Diseases                           | Extended     | 339,668   |
| Integrative & Complementary Medicine          | Extended     | 72,228    |
| Integrative & Complementary Medicine          | Traditional  | 72,228    |
| International Relations                       | Traditional  | 162,992   |
| International Relations                       | Extended     | 165,720   |
| Law                                           | Traditional  | 226,907   |
| Legal Medicine                                | Extended     | 40,280    |
| Limnology                                     | Traditional  | 39,672    |
| Linguistics                                   | Traditional  | 156,533   |
| Linguistics                                   | Extended     | 273,025   |
| Literary Reviews                             | Traditional  | 161,784   |
| Literary Theory & Criticism                   | Traditional  | 92,415    |
| Literary Theory Criticism                     | Traditional  | 3         |
| Literature                                    | Traditional  | 181,859   |
| Literature                                    | Extended     | 646,905   |
Table B.2: We exclude from inference the WoS categories listed below whose contents are not likely to overlap with arXiv’s (continued).

| Category                                      | Typology | Count   |
|-----------------------------------------------|----------|---------|
| Literature American                           | Traditional | 12      |
| Literature, African, Australian, Canadian     | Traditional | 17,278  |
| Literature, American                          | Traditional | 26,529  |
| Literature, British Isles                     | Traditional | 33,078  |
| Literature, German, Dutch, Scandinavian       | Traditional | 29,165  |
| Literature, Romance                           | Traditional | 104,049 |
| Literature, Slavic                            | Traditional | 15,173  |
| Management                                    | Traditional | 365,430 |
| Marine & Freshwater Biology                   | Traditional | 215,811 |
| Marine & Freshwater Biology                   | Extended  | 247,068 |
| Medical Ethics                                | Extended  | 26,182  |
| Medical Ethics                                | Traditional | 26,182  |
| Medical Informatics                           | Extended  | 77,036  |
| Medical Informatics                           | Traditional | 77,036  |
| Medical Laboratory Technology                 | Traditional | 91,106  |
| Medical Laboratory Technology                 | Extended  | 93,651  |
| Medicine General Internal                     | Traditional | 794,946 |
| Medicine, General & Internal                  | Traditional | 907,739 |
| Medicine, Legal                               | Traditional | 40,280  |
| Medicine, Research & Experimental             | Traditional | 543,200 |
| Medieval & Renaissance Studies                | Traditional | 72,264  |
| Metallurgy & Metallurgical Engineering        | Traditional | 362,643 |
| Metallurgy & Metallurgical Engineering        | Extended  | 539,933 |
| Metallurgy Metallurgical Engineering          | Traditional | 164,388 |
| Mineralogy                                    | Traditional | 123,996 |
| Mineralogy                                    | Extended  | 125,890 |
| Mining & Mineral Processing                   | Traditional | 64,704  |
| Mining & Mineral Processing                   | Extended  | 97,881  |
| Mining Mineral Processing                     | Traditional | 27,357  |
| Multidisciplinary Sciences                    | Traditional | 864,971 |
| Music                                         | Extended  | 168,745 |
| Music                                         | Traditional | 168,745 |
| Mycology                                      | Traditional | 47,602  |
| Mycology                                      | Extended  | 47,823  |
| Nursing                                       | Extended  | 197,697 |
| Nursing                                       | Traditional | 197,697 |
| Nutrition & Dietetics                         | Traditional | 252,581 |
| Nutrition & Dietetics                         | Extended  | 253,652 |
| Obstetrics & Gynecology                       | Traditional | 389,689 |
| Obstetrics & Gynecology                       | Extended  | 427,290 |
| Obstetrics Gynecology                         | Traditional | 32,525  |
| Oceanography                                  | Traditional | 157,749 |
| Oceanography                                  | Extended  | 160,252 |
| Oncology                                      | Traditional | 1,404,745 |
| Oncology                                      | Extended  | 1,408,017 |
| Operations Research & Management Science      | Traditional | 249,652 |
| Operations Research & Management Science      | Extended  | 252,977 |
| Ophthalmology                                 | Traditional | 335,794 |
| Ophthalmology                                 | Extended  | 336,700 |
| Ornithology                                   | Traditional | 26,882  |
| Orthopedics                                   | Traditional | 242,782 |
| Orthopedics                                   | Extended  | 244,714 |
| Otorhinolaryngology                           | Traditional | 135,459 |
| Otorhinolaryngology                           | Extended  | 137,217 |
| Paleontology                                  | Traditional | 62,404  |
Table B.2: We exclude from inference the WoS categories listed below whose contents are not likely to overlap with arXiv’s (continued).

| Category                                      | Typology  | Count  |
|-----------------------------------------------|-----------|--------|
| Paleontology                                  | Extended  | 62,791 |
| Parasitology                                  | Traditional | 98,060 |
| Parasitology                                  | Extended  | 100,012|
| Pathology                                     | Traditional | 307,609|
| Pathology                                     | Extended  | 309,306|
| Pediatrics                                    | Traditional | 473,447|
| Pediatrics                                    | Extended  | 482,052|
| Peripheral Vascular Disease                   | Traditional | 453,394|
| Pharmacology & Pharmacy                       | Traditional | 967,993|
| Pharmacology & Pharmacy                       | Extended  | 1,274,526|
| Pharmacology Pharmacy                         | Traditional | 137,791|
| Philosophy                                    | Traditional | 230,405|
| Philosophy                                    | Extended  | 237,410|
| Physical Geography                            | Extended  | 114,415|
| Physiology                                    | Traditional | 287,400|
| Physiology                                    | Extended  | 291,511|
| Poetry                                        | Traditional | 41,586 |
| Political Science                             | Traditional | 382,117|
| Primary Health Care                           | Traditional | 52,820 |
| Psychiatry                                    | Traditional | 586,368|
| Psychiatry                                    | Extended  | 593,529|
| Psychology                                    | Traditional | 212,547|
| Psychology                                    | Extended  | 1,020,210|
| Psychology Multidisciplinary                  | Traditional | 11,083 |
| Psychology, Applied                           | Traditional | 101,369|
| Psychology, Biological                        | Traditional | 55,700 |
| Psychology, Clinical                          | Traditional | 185,894|
| Psychology, Developmental                     | Traditional | 107,005|
| Psychology, Educational                       | Traditional | 52,964 |
| Psychology, Experimental                      | Traditional | 169,345|
| Psychology, Mathematical                      | Traditional | 13,123 |
| Psychology, Multidisciplinary                 | Traditional | 265,542|
| Psychology, Psychoanalysis                    | Traditional | 26,121 |
| Psychology, Social                            | Traditional | 82,545 |
| Public Administration                          | Traditional | 73,580 |
| Public Administration                          | Extended  | 207,916|
| Public Environmental Occupational Health       | Traditional | 77,977 |
| Public, Environmental & Occupational Health   | Traditional | 698,326|
| Public, Environmental & Occupational Health   | Extended  | 786,526|
| Radiology, Nuclear Medicine & Medical Imaging | Traditional | 636,954|
| Radiology, Nuclear Medicine & Medical Imaging | Extended  | 638,708|
| Regional & Urban Planning                     | Traditional | 134,920|
| Rehabilitation                                | Traditional | 178,866|
| Rehabilitation                                | Extended  | 180,271|
| Religion                                      | Extended  | 270,286|
| Religion                                      | Traditional | 270,286|
| Reproductive Biology                          | Traditional | 169,017|
| Reproductive Biology                          | Extended  | 171,394|
| Research & Experimental Medicine              | Extended  | 547,770|
| Respiratory System                            | Traditional | 396,574|
| Respiratory System                            | Extended  | 397,705|
| Rheumatology                                  | Traditional | 220,643|
| Rheumatology                                  | Extended  | 221,803|
| Social Issues                                 | Traditional | 87,334 |
| Social Issues                                 | Extended  | 89,867 |
Table B.2: We exclude from inference the WoS categories listed below whose contents are not likely to overlap with arXiv’s (continued).

| Category                                      | Typology     | Count    |
|-----------------------------------------------|--------------|----------|
| Social Sciences - Other Topics                | Extended     | 483,051  |
| Social Sciences Interdisciplinary             | Traditional  | 906      |
| Social Sciences Mathematical Methods          | Traditional  | 296      |
| Social Sciences, Biomedical                   | Traditional  | 83,498   |
| Social Sciences, Interdisciplinary            | Traditional  | 247,255  |
| Social Sciences, Mathematical Methods         | Traditional  | 60,023   |
| Social Work                                   | Traditional  | 62,286   |
| Social Work                                   | Extended     | 62,505   |
| Sociology                                     | Traditional  | 225,233  |
| Sociology                                     | Extended     | 232,336  |
| Soil Science                                  | Traditional  | 93,640   |
| Sport Sciences                                | Traditional  | 238,951  |
| Sport Sciences                                | Extended     | 240,200  |
| Substance Abuse                               | Traditional  | 100,091  |
| Substance Abuse                               | Extended     | 100,783  |
| Surgery                                       | Traditional  | 1,150,579|
| Surgery                                       | Extended     | 1,159,981|
| Theater                                       | Extended     | 50,387   |
| Theater                                       | Traditional  | 50,387   |
| Toxicology                                    | Traditional  | 274,198  |
| Toxicology                                    | Extended     | 274,890  |
| Transplantation                               | Traditional  | 245,611  |
| Transplantation                               | Extended     | 246,998  |
| Transportation                                | Traditional  | 143,431  |
| Transportation                                | Extended     | 251,724  |
| Transportation Science & Technology           | Traditional  | 141,134  |
| Tropical Medicine                            | Extended     | 87,544   |
| Tropical Medicine                            | Traditional  | 87,544   |
| Urban Studies                                 | Traditional  | 85,644   |
| Urban Studies                                 | Extended     | 88,481   |
| Urology & Nephrology                          | Traditional  | 415,848  |
| Urology & Nephrology                          | Extended     | 420,386  |
| Urology Nephrology                            | Traditional  | 24        |
| Veterinary Sciences                           | Traditional  | 362,412  |
| Veterinary Sciences                           | Extended     | 365,769  |
| Virology                                      | Traditional  | 152,546  |
| Virology                                      | Extended     | 153,859  |
| Water Resources                               | Traditional  | 274,775  |
| Water Resources                               | Extended     | 275,614  |
| Women’s Studies                               | Extended     | 75,124   |
| Women’s Studies                               | Traditional  | 75,124   |
Table C.1: Keyword performance in the complete arXiv corpus is highest in 2019. Its decline in earlier years suggests the need for continuous maintenance of term lists. Scores are given for the positive class and the support column refers to the number of AI-relevant articles.

| Year | Precision | Recall | \( F_1 \) | Support | Total  |
|------|-----------|--------|-----------|---------|--------|
| 2010 | .50       | .27    | .35       | 1,379   | 70,286 |
| 2011 | .54       | .24    | .33       | 2,025   | 76,605 |
| 2012 | .63       | .25    | .36       | 3,370   | 84,389 |
| 2013 | .65       | .25    | .36       | 4,561   | 92,866 |
| 2014 | .66       | .31    | .43       | 4,896   | 97,598 |
| 2015 | .71       | .36    | .48       | 6,663   | 105,128|
| 2016 | .78       | .41    | .54       | 10,566  | 113,436|
| 2017 | .77       | .44    | .56       | 15,670  | 123,781|
| 2018 | .77       | .48    | .59       | 23,891  | 140,392|
| 2019 | .80       | .49    | .61       | 34,359  | 155,840|
| All  | .76       | .43    | .55       | 103,380 | 1,060,321|

Table C.2: In evaluation against arXiv test data, the keyword-classifier hybrid developed by Elsevier shows improvements over our baseline keyword solution. Longitudinal imbalance in the training data results in higher performance in recent years.

| Year | Positive Class | Negative Class | Wtd. Avg. |
|------|----------------|----------------|-----------|
|      | Precision      | Recall         | \( F_1 \) | Support  | Precision | Recall | \( F_1 \) | Support  | F | Support |
| 2010 | .50            | .31            | .38       | 138      | .99       | .99    | .99       | 6,891    | .98 | 7,029  |
| 2011 | .50            | .30            | .38       | 202      | .98       | .99    | .99       | 7,458    | .97 | 7,660  |
| 2012 | .58            | .26            | .36       | 337      | .97       | .99    | .98       | 8,102    | .96 | 8,439  |
| 2013 | .60            | .28            | .39       | 456      | .96       | .99    | .98       | 8,831    | .95 | 9,287  |
| 2014 | .59            | .31            | .41       | 489      | .96       | .99    | .98       | 9,271    | .95 | 9,760  |
| 2015 | .69            | .42            | .52       | 666      | .96       | .99    | .97       | 9,847    | .95 | 10,513 |
| 2016 | .75            | .45            | .57       | 1,057    | .95       | .98    | .97       | 10,287   | .93 | 11,344 |
| 2017 | .74            | .49            | .59       | 1,567    | .93       | .98    | .95       | 10,811   | .91 | 12,378 |
| 2018 | .75            | .55            | .64       | 2,389    | .91       | .96    | .94       | 11,650   | .89 | 14,039 |
| 2019 | .81            | .55            | .66       | 3,436    | .88       | .96    | .92       | 12,148   | .86 | 15,584 |
| All  | .74            | .49            | .59       | 10,737   | .94       | .98    | .96       | 95,296   | .92 | 106,033|
Table C.3: All-subjects SciBERT test performance is highest in recent years, primarily because of longitudinal imbalance in the training data. There is also a class imbalance of about 9:1 in favor of negative examples whose effect on performance is apparent.

| Year | Precision | Recall | \( F_1 \) | Support | Precision | Recall | \( F_1 \) | Support | \( F_1 \) | Support |
|------|------------|--------|------------|---------|-----------|--------|------------|---------|------------|---------|
| 2010 | .64        | .63    | .64        | 138     | .99       | .99    | .99        | 6,891   | .99        | 7,029   |
| 2011 | .65        | .63    | .64        | 202     | .99       | .99    | .99        | 7,458   | .98        | 7,660   |
| 2012 | .74        | .69    | .72        | 337     | .99       | .99    | .99        | 8,102   | .98        | 8,439   |
| 2013 | .80        | .74    | .77        | 456     | .99       | .99    | .99        | 8,831   | .98        | 9,287   |
| 2014 | .74        | .73    | .74        | 489     | .99       | .99    | .99        | 9,271   | .97        | 9,760   |
| 2015 | .78        | .79    | .78        | 666     | .99       | .98    | .99        | 9,847   | .97        | 10,513  |
| 2016 | .82        | .84    | .83        | 1,057   | .98       | .98    | .98        | 10,287  | .97        | 11,344  |
| 2017 | .83        | .89    | .85        | 1,567   | .98       | .97    | .98        | 10,811  | .96        | 12,378  |
| 2018 | .83        | .90    | .87        | 2,389   | .98       | .96    | .97        | 11,650  | .95        | 14,039  |
| 2019 | .87        | .89    | .88        | 3,436   | .97       | .96    | .97        | 12,148  | .95        | 15,584  |
| All  | .83        | .85    | .84        | 10,737  | .98       | .98    | .98        | 95,296  | .97        | 106,033 |