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

June 7, 2020

ABSTRACT

We describe a strategy for identifying the universe of research publications relevant to the application and development of artificial intelligence. The approach leverages the arXiv corpus of scientific preprints, in which authors choose subject tags for their papers from a set defined by editors. We compose a functional definition of AI relevance by learning these subjects from paper metadata, and then inferring the arXiv-subject labels of papers in larger corpora: Clarivate Web of Science, Digital Science Dimensions, and Microsoft Academic Graph. This yields predictive classification F_1 scores between .75 and .86 for Natural Language Processing (`cs.CL`), Computer Vision (`cs.CV`), and Robotics (`cs.RO`). For a single model that learns these and four other AI-relevant subjects (`cs.AI`, `cs.LG`, `stat.ML`, and `cs.MA`), we see precision of .83 and recall of .85. We evaluate the out-of-domain performance of our classifiers against other sources of topic information and predictions from alternative methods. We find that a supervised solution can generalize to identify publications that belong to the high-level fields of study represented on arXiv. 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.