Detecting dubious research with SciBERT

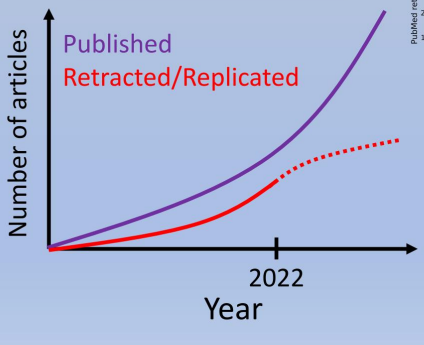
Eric Sun
Department of Biomedical Data Science, Stanford University

Stanford University

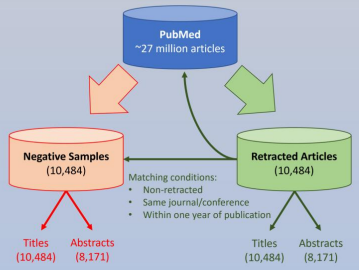
Problem

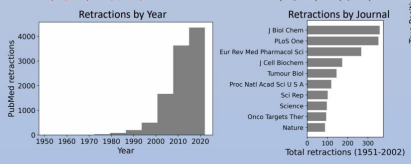
Over the past few decades, the rate of scientific publication has increased dramatically. At the same time, the number of retractions and failed replications of studies have also increased.[1] This replication crisis across many scientific disciplines has heightened scrutiny of scientific papers and their reported results.[2] However, the incentive for researchers to pursue replication of prior research is limited and are typically outweighed by the costs. As such, there is a need for computational methods for prioritizing replication efforts on scientific findings that are most dubious (i.e. a high estimated probability of retraction) and highest impact (i.e. a large number of citations). The goal of this final project is to address the first desiderata by building a classifier for detecting dubious research articles from their titles and abstracts.

Background



Data





Methods

SciBERT transformer model

- 1.14 million full-text research papers
- 20% CS + 80% Biomedical
- Specially tailored SciVocab

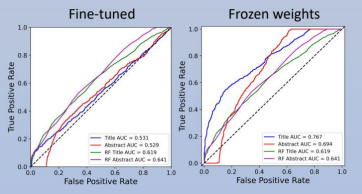
Random forest baseline

- Trained on TF-IDF vectors

Experiments

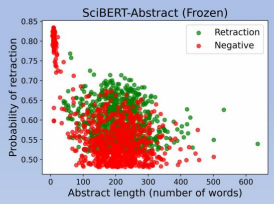
Fine-tuning: Training updates to all SciBERT weights
Frozen weights: Training updates only to classifier head

Parameters: 20 epochs with early stopping; optimizer = SGD; loss = BCE, lr=2e-5; decay = 0.01; batch=16

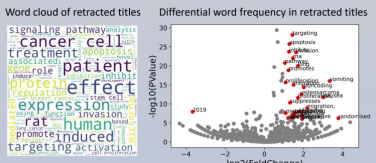


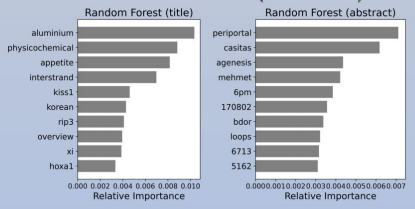
Model	AUROC (test)	F1 Score (test)
SciBERT-Title (frozen)	**0.767**	0.694
SciBERT-Abstract (frozen)	0.694	**0.713**
SciBERT-Title (finetune)	0.531	0.665
SciBERT-Abstract (finetune)	0.529	0.669
Random Forest-Title	0.619	0.601
Random Forest-Abstract	0.641	0.569

Table 1: Evaluation metrics reported for the test set (10% of total data) for the title-trained and abstract-trained SciBERT models and baseline random forest models



Analysis





Conclusions

- Machine learning models can classify retracted papers from matched negative samples using title and abstract text.
- Including additional features may increase performance but could also introduce bias.
- In the future, it is necessary to evaluate the efficacy of the predicted probability of retraction (in conjunction with scientific impact metrics) for prioritizing replication efforts.

1. Loken, E., Gelman, A., 2017. Measurement error and the replication crisis. Science 355, 584–585. https://doi.org/10.1126/science.aal3618
2. Ioannidis, J.P.A., 2005. Why Most Published Research Findings Are False. PLOS Medicine 2, e124. https://doi.org/10.1371/journal.pmed.0020124
3. Beltagy, I., Lo, K., Cohan, A., 2019. SciBERT: A Pretrained Language Model for Scientific Text. arXiv:1903.10676 [cs].