Detecting dubious research with SciBERT

Stanford CS224N Custom Project

Eric Sun
Department of Biomedical Data Science
Stanford University
edsun@stanford.edu

Abstract

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. Using the PubMed database, I retrieved over 10,000 titles and abstracts of retracted scientific papers along with matched negative controls. I finetuned a SciBERT model [3] to classify retracted papers from negative controls with only the titles or abstracts. Under finetuning with frozen SciBERT weights, these models outperformed a benchmark random forest classifier model trained on token frequency vectors. In combination with other metrics such as the number of citations, these models may be useful in prioritizing certain research articles to be investigated more thoroughly via replication efforts, which would be beneficial for the entire scientific ecosystem.

1 Key Information to include

- Mentor: Benjamin Newman
- External Collaborators (if you have any): No
- Sharing project: STAT 211

2 Introduction

Over the past several decades, there has been an exponential increase in the number of publications that are being peer-reviewed and added to scientific repositories. As a direct result of the increasing number of published papers, the number of retractions are also increasing steadily [1]. However, given the pace of increase in publication rate, it is unclear if important safeguards for the integrity of the scientific literature like replication efforts and retractions can keep up. In some disciplines like psychology and medicine, there are additional barriers like the high cost of recruiting a participant cohort, matching demographics across samples, and abiding by rigorous ethical standards and intellectual property rights that compound the difficulty in replicating findings and identifying dubious papers for retraction [4]. Altogether, this “replication crisis" has become a concerning trend in some fields that has decreased confidence in published results [2].

Despite mounting concerns surrounding the scientific literature, there are few incentives in place for researchers to undertake replication efforts in lieu of original research [5]. Furthermore, there are many barriers to replication studies including lack of resources available to recapitulate the same experimental conditions, lack of detailed documentation of the experiment, and that negative findings are not evidence that the original finding's were wrong [5]. As a result, the number of replication studies are relatively low and there are likely many dubious published findings that have not been retracted. With the publication rate looking to increase exponentially in the foreseeable future, there is a dire need for designing reliable and high-throughput methods to screen new scientific publications and prioritize the highest-risk publications for spending our limited replication efforts on (see Fig. 1A for a graphical illustration). In particular, it would be important for these methods to identify papers that are (1) flawed or likely to be retracted, and (2) likely to have a large impact on the scientific discipline. The first desiderata concerns the papers with highest likelihood of including false findings. The second desiderata prioritizes the papers that, if not retracted, would have a large deleterious impact on scientific progress. The second desiderata is fairly easy to estimate with standard scientific impact metrics like citation counts or citation rates. In this project, I focus on designing a model for flagging dubious research to fulfill the first desiderata.

My approach to building a classifier for dubious research is to use only features that can be derived from either the title or abstract of the research article to compute a probability of retraction for that article. In order to achieve this goal, I curated a set of titles from over 10,000 retracted research articles and a subset of abstracts from over 8,000 retracted research articles from the PubMed repository [6]. In parallel, I also matched each of these retracted articles with a negative sample article that was non-retracted, published within one year of the positive sample, and published in the same journal as the positive sample. Then, I trained several finetuned variations of the SciBERT transformer model on the SciVocab-tokenized titles or abstracts of these research articles to classify retractions from negative samples. These models were benchmarked against a random forest model that was trained to classify retractions from word frequency vectors. Furthermore, I investigated the important features of the random forest model and SciBERT models in classifying retractions.

3 Related Work

The only prior work on classifying retractions was a thesis recently published at Penn State University [7], which leveraged a random forest model to classify retracted papers in the Retraction Watch dataset. In the random forest model, the authors included a large set of diverse features including different text and sentence embeddings of the abstracts, word frequencies, publication metadata, and metadata on the authors (e.g. h-index of authors, frequency of publication, names, etc). The best model reported test accuracy of 73.32 and F1 score of 71.77. Unfortunately, due to the lengthy approval process for accessing Retraction Watch data, I was unable to use that data in training my models and instead curated my own dataset. As such, direct comparisons in performance between my model and the aforementioned random forest model are impossible.

A primary motivation of my project is to simplify the feature set to only use the text of the article instead of other metadata. This is because these other metadata, such as those pertaining to the authors or journals, are likely to directly introduce bias into the algorithm (e.g. authors with non-English names penalized unfairly, authors from low-resource universities penalized unfairly, low-ranked journals penalized unfairly, etc). This is important because in order for such a model to be useful in practice, minimal bias against any particular demographic of authors or journals is desirable. As such, to compare my model to a baseline, I independently trained a random forest model on word frequency vectors, which were the most important text-related feature in the related work [7].

4 Approach

In this final project, I am training a SciBERT-based classifier to predict whether the title or abstract of a given research article is from a retracted article (as a proxy for dubious research) or a negative (non-retracted) control article. The inputs to the text classifier model are SciVocab-tokenized versions of either the titles or the abstracts (capped at a maximum token length of 512). The output of the model for each of the input samples is a probability vector of length 2 that is computed from the softmax operation and specifies the probability of retraction (first component) and non-retraction (second component), with the total probability summing to unity. The probability of retraction can be interpreted as how dubious the research article is and could potentially be used to prioritize the replication efforts towards articles with higher probabilities of retraction.

SciBERT was developed according to the original BERT architecture, a bidrectional Transformer that is trained on two tasks: (1) predicting randomly masked tokens, and (2) predicting whether two sentences follow each other [3]. Instead of training the model on the BERT corpora (Wikipedia and news articles), SciBERT was trained on scientific text. The scientific text corpora was obtained as a random sample of 1.14 million full-text research papers retrieved from Semantic Scholar. The coarse breakdown of the papers is approximately 18% from computer science and 82% from biomedical research. In total, there were 3.17 billion tokens in the corpora. Using this corpora, SciVocab was constructed as a 30,000-word vocabulary using the same approach as for BERT. The overlap between SciVocab and BERT vocabulary was only 42%, which highlights the significant differences in language usage between the two corpora [3]. I used a modified SciBERT model that included a final classification head, and explored two methods of training the model: (1) finetuning all weights across both the classifier head and the SciBERT model, and (2) freezing the weights of the SciBERT model and only tuning the weights of the classifier head.

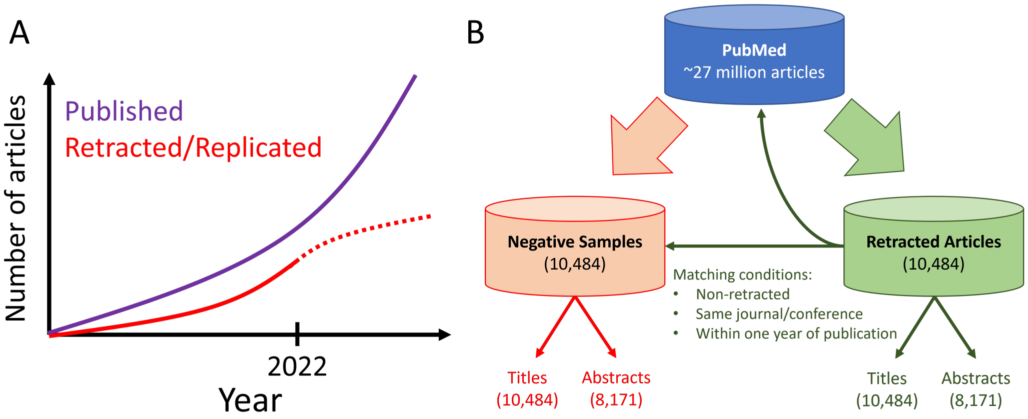

Figure 1: (a) Schematic representation of the problem addressed in this final project: exponentially increasing publication rate that is saturating the limited resources available for replication/retraction efforts, motivating optimal allocation of these resources. (b) Schematic representation of the data curation pipeline used in this study where I used the PubMed eSearch API to query retracted articles, parse their title and abstract data, and match to negative samples based on associated metadata (journal, year of publication).

5 Experiments

5.1 Data

Originally, I had planned to use the Retraction Watch dataset, which is freely available for academic uses after obtaining approval and signing a contract. However, given the demanding approval process of accessing Retraction Watch data (e.g. need for a faculty sponsor and Stanford administrator to sign a contract) and the time constraints for this final project, I opted instead to curate my own dataset, which I plan to make publicly accessible. To do this, I wrote a pipeline to access PubMed articles through its eSearch API, filter for papers tagged with a retraction notice, parse their title and abstract fields, filter out low-quality hits, and retrieve relevant metadata such as year and venue of publication. After retrieving these positive samples of retracted articles, I used the metadata to match each retracted article with a negative sample that was published within one year of the retracted article and in the same journal/venue. In total, I recovered data from 10,484 retracted articles (20,968 total articles after including matched negative samples), which were published in 2518 unique journals and spanned several decades from 1951-2022. Interestingly, the rate of retractions appears to increase exponentially with time up to the current time point (Fig. 2a), which supports the motivating principle of this work as highlighted in Fig. 1a. Another interesting trend that appears in the data is an over-representation of biomedical journals, particularly high-impact journals, among the retractions, which may point to the higher level of scrutiny and susceptibility to fraudulent results under these conditions.

When querying for just abstracts, some of the retracted articles did not have associated abstracts (i.e. just a retraction notice of the title) and others with very short lengths were filtered out. In total, I retrieved 8,171 full abstracts (16,342 total abstracts after including matched negative samples). The abstracts were tokenized and truncated to a maximum length of 512 tokens according to SciVocab.

5.2 Evaluation method

I split my data into train (80%), development (10%), and test (10%) sets by randomly assigning the paired retracted articles and negative samples to each of these sets. To evaluate model performance, I am using AUROC and F1 score (described in greater detail in the proposal). I will benchmark my model against a random forest classifier that uses textual features like bag-of-words (BOW) to discriminate retracted articles from negative samples, which is the same model that was used in a previous work [7]. Furthermore, I verified the consistency of these evaluation metrics under both several resampling iterations of the negative samples; different splits of the data into training, development, and test sets; and under bootstrap sampling of the test sets. Due to the low number of re-samplings done in each case, I only show the point metrics for the first set of results instead of reporting a confidence interval.

5.3 Experimental details

To build the model, I am finetuning the SciBERT (uncased) model that was pretrained on SciVocab and is available in HuggingFace AllenAI: `https://huggingface.co/allenai/scibert_scivocab_uncased`. The SciBERT model was pre-trained on a large scientific corpus of computer science and biomedical texts and uses the standard BERT architecture [3]. Details on SciBERT are available in the previous Approach section. I implemented the model with a classifier head via the “AutoModelForSequenceClassification" method of HuggingFace, which essentially replaces the final layer with a classification head. In this case, I am using num_labels=2 (binary classification). I have written the code to split the data into train, development, and test sets and format them correctly for the HuggingFace API. For sequences with token length less than the maximum length (512), I used padding for the remaining tokens. I performed model finetuning using the HuggingFace API with binary cross entropy (BCE) as the loss function and a learning rate of 2×10^{-5}, weight decay of 0.01, and batch size of 16. These hyperparameters were set heuristically after experimenting with other values that were a few orders of magnitude larger or smaller. During training, the loss and accuracy for the train and development sets were printed after each epoch. I explored two setups for model training. In the first setup, back-propagation of the error to update all weights (SciBERT and classifier head) was done. This is referred to as “fine-tuned" in this project report. In the second setup, referred to as “frozen weights" in this project report, back-propagation of the error was only done to update the weights of the classifier head, leaving the SciBERT pre-trained weights frozen. Training for the title text model took around 2 hours while training of the abstract text model took around 4 hours on the CS 224N Microsoft Azure virtual machine. The best model as determined by minimum loss on the development set was save after 20 epochs had elapsed.

5.4 Results

After curating a dataset of titles and abstracts from retracted articles and matched negative controls, I characterized relevant properties of the dataset, which includes examining the annual number of retractions across time (Fig. 2a), the journals with the greatest number of retractions (Fig. 2b), the words that are most common among retracted articles and negative samples (Fig. 1c), and the words that have statistically significant differences in frequency between retracted articles and negative controls (using Fisher’s exact test, Fig. 2d). Interestingly, there are a significantly greater number of words that are overrepresented in retracted articles than underrepresented in retracted articles, which suggests that a reliable signal for detecting dubious scientific articles may be present in the inclusion of certain words rather than the exclusion of others. The identification of word frequency differences between the retracted article titles and negative sample titles also hints at the potential to classify retracted articles from negative samples using text alone.

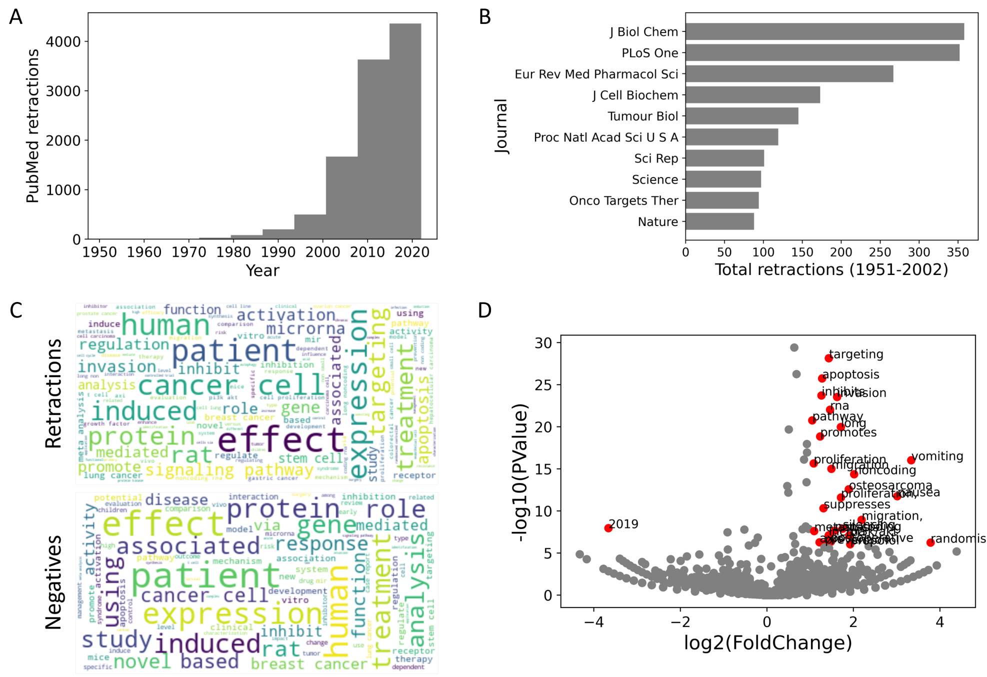

Figure 2: Properties of retracted article titles in PubMed. (a) Annual retractions has increased exponentially. (b) Ranking of journals with the greatest number of retractions. (c) Word clouds of most frequent words in the positive samples (retracted titles) and negative samples. (d) Volcano plot of differential word frequency analysis of words between positive and negative samples (labeled red words correspond to Bonferroni-adjusted p-value of 0.05 and log2 fold change greater than one, Fisher's exact test).

In order to set a baseline performance for a retraction classifier, I trained two independent random forest models to classify between retracted article titles or abstracts and negative sample titles or abstracts from word frequency vectors. The word frequency vectors were computed with term frequency–inverse document frequency (TF-IDF), which reflects the importance of a word to a document (in our case, either the title or abstract) by normalizing the raw word count in a document by the number of documents containing that count in the corpus. I used the same train, development, and test sets described in the Evaluation Method section to train and evaluate the random forest classifier models. In line with the results from the preliminary statistical analyses, the random forest baseline classifiers were able to discriminate retracted titles (AUROC = 0.619, F1 score = 0.601) and retracted abstracts (AUROC = 0.641, F1 score = 0.569). There was no noticeable difference in performance between the random forest baseline models trained on titles or abstracts. This is shown in Table 1.

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

Next, I implemented and finetuned all weights in the SciBERT classifier on the same training set as used for the random forest baseline models. I trained separate models on abstracts and titles.

The models were trained for 20 epochs and the best model (as determined by validation loss on the development set) was retrieved and evaluated on the test set. Surprisingly, the models did not outperform the random forest baseline models (see Table 1), which is illustrated in Fig. 3b, which was not what I had expected. There was also a notable drop in accuracy from the train set to the test set of around 10-15%, which suggested that the models may have been overfit on the training dataset. The increase of validation loss after the first epoch of training also suggests that the finetuned SciBERT models were susceptible to overfitting (Fig. 3a). As such, I designed a secondary training setup where the SciBERT weights are frozen and only the classifier head weights are updated. For the frozen SciBERT models, there was significantly better performance than baseline for discriminating retracted titles (AUROC = 0.767, F1 score = 0.694) and for discriminating retracted abstracts (AUROC = 0.694, F1 score = 0.713), which is illustrated in Fig. 3c. In the case of all models, there was not a significant trend of better performance with abstract text or title text (see Table 1).

Overall, the performance of these models was better than I expected because intuitively, the task of classifying retracted articles is very difficult and something that even well-trained scientists cannot typically do by reading the title or abstract of a paper alone. As a result, further work using the approach of building SciBERT-based models would likely lead to more reliable classifiers that may ultimately become practically useful for prioritizing resources for replication studies and preserving scientific integrity.

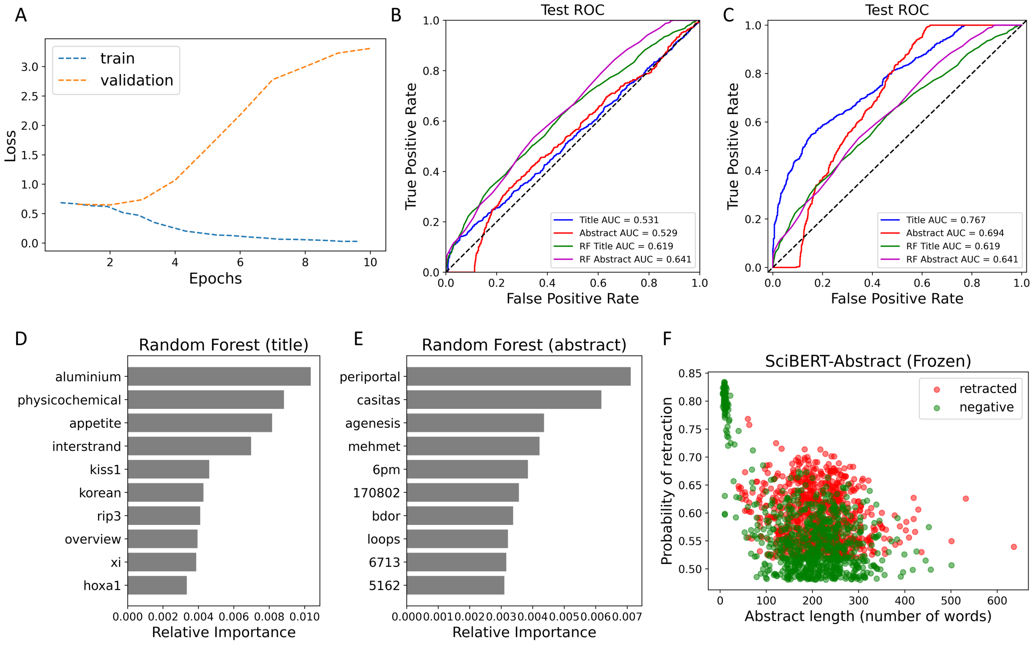

Figure 3: Evaluation and interpretation of trained models. (a) Train and evaluation loss on a dev set for the finetuned SciBERT-Title model. (b) Receiver operator characteristic (ROC) curves for the finetuned SciBERT-Title, finetuned SciBERT-Abstract, and random forest baseline models. (c) Receiver operator characteristic (ROC) curves for the frozen weight SciBERT-Title, frozen weight SciBERT-Abstract, and random forest baseline models. (d) Feature importance ranking for the random forest baseline model trained on article title. (e) Feature importance ranking for the random forest baseline model trained on article abstract. (f) Predicted probability of retraction from the frozen weight SciBERT-Abstract model as a function of abstract length.

6 Analysis

In order to understand the predictions made by the SciBERT classifier models, I employed two orthogonal approaches. The first was to interpret the most important features (i.e. tokens) in the random forest baseline models as a proxy for important discriminative features for predicting retraction texts. Feature importance is easy to compute in the random forest architecture due to the transparent nature of decision trees (i.e. they split on features in order of importance). The ranking of the top ten tokens are shown in Fig. 3d for the title-based model and in Fig. 3e for the abstract-based model. It is important to note that the top features used for classification are different from the most statistically different features recovered from the analysis in Fig. 2d, which suggests that successful classifier models are leveraging information outside of independent word frequencies in order to classify retracted texts. The second approach that I took to analyze the SciBERT models was to correlate the predictions against different metadata in hopes of uncovering some associated features that the SciBERT model may be indirectly measuring. A particular characteristic that stood out in the SciBERT-Abstract model was the length of the input. Abstracts of very small lengths (top left cluster in Fig. 3f) among the negative samples were virtually all misclassified as retracted texts. As such, in the future, improved model performance may be achieved by filtering out these outlier examples during both training and testing.

7 Conclusion

In this final project, several learning goals that I achieved was on understanding the BERT architecture through SciBERT, implementing and training a text classification model, and integrating traditional word frequency approaches (e.g. TF-IDF vectors) with deep learning models. The main achievements of this final project were: (1) curation of a novel dataset of more than 10,000 retracted titles and over 8,000 retracted abstracts along with a matched set of negative samples, which can be made publicly available for other researchers; (2) demonstration that text classification models can successfully discriminate retracted from non-retracted texts (i.e. there is some signal of retraction within the text itself!); and (3) the first deep learning, SciBERT-based retraction classifiers that outperform random forest baseline models, which are comparable in architecture to previous works. Importantly, these models only rely on the text data and as a result, will be free from some biases that constrain the application of previous classifiers that use information on the authors and journals.

In the future, it will be important to consider the efficacy of these SciBERT-based classifiers in optimizing the use of resources for replication efforts when considered in tandem with impact metrics like citation counts. As stated previously, adding non-textual features may increase the bias of the model, but future work in investigating the tradeoff between potential bias and increased performance when including features like author and journal metadata may be insightful. Another avenue that would be worth exploring is comparing the current models to a CNN model, which would not rely on pre-training, but likely be better suited for text classification than RNNs or simple linear neural networks. This may provide a baseline approach to understand what level of performance can be achieved without pre-training on millions of scientific documents.

References

[1] Eric Loken and Andrew Gelman. Measurement error and the replication crisis. *Science*, 355(6325):584–585, February 2017. Publisher: American Association for the Advancement of Science.

[2] John P. A. Ioannidis. Why Most Published Research Findings Are False. *PLOS Medicine*, 2(8):e124, August 2005. Publisher: Public Library of Science.

[3] Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A Pretrained Language Model for Scientific Text. *arXiv:1903.10676 [cs]*, September 2019. arXiv: 1903.10676.

[4] Scott O. Lilienfeld. Psychology's Replication Crisis and the Grant Culture: Righting the Ship. *Perspectives on Psychological Science*, 12(4):660–664, July 2017. Publisher: SAGE Publications Inc.

[5] Gerd Gigerenzer. Statistical Rituals: The Replication Delusion and How We Got There. *Advances in Methods and Practices in Psychological Science*, 1(2):198–218, June 2018. Publisher: SAGE Publications Inc.

[6] Richard J. Roberts. PubMed Central: The GenBank of the published literature. *Proceedings of the National Academy of Sciences*, 98(2):381–382, January 2001. Publisher: Proceedings of the National Academy of Sciences.

[7] Sai Ajay Modukuri, Sarah Rajtmajer, Anna Cinzia Squicciarini, Jian Wu, and C Lee Giles. Understanding and Predicting Retractions. page 8.