The goal of this project is to try and use natural language processing to provide some amount of insight into the quality of COVID-19 related preprints (taken from medRxiv and bioRxiv).

**Problem statement**: COVID-19 is the most significant pandemic to occur during the modern scientific era, and large sections of the scientific establishment have rapidly pivoted their efforts to better understand this disease and the virus that causes it. This exacerbates a pre-existing problem in modern science communication as the amount of new research being published every day in a given field is much more than any one person can conceivably read. This is especially problematic when work has not yet undergone peer review, as the peer review process provides a screen to separate research into higher and lower quality journals (though of course this process is imperfect) and catches errors or inaccuracies. This is a problem for researchers, because it makes it difficult to keep up with the most recent work in other labs without wasting time on low quality research (the quality of which is often not apparent until significant time has already been invested). It is also a problem for policy makers, as the firehose of information makes it less likely that they are accessing the most accurate sources at any given time and they often do not have the domain expertise to differentiate low and high quality work in every field. This problem can be mitigated by developing a solution to rapidly rate scientific papers for quality, which will allow users to more quickly parse the deluge of information coming out on a daily basis without investing time reading likely low-quality work.

**Strategy**: Implement an end to end machine learning solution, which will take as input the text of a specific preprint paper and output a ranking between 0 and 100% in terms of estimated quality. Quality in this case is an imperfect measure, and will be defined using an unsupervised approach by quantifying similarity to papers which have undergone peer review and been published in ‘respectable’ scientific journals, as compared with similarity to papers which have been published in ‘predatory’ journals.

**Definitions**:

Respectable journals – Those which have a verified peer review process, are tied to an institution, and enforce common practices in research integrity (not re-publishing existing articles, data integrity, etc). For this project I will

Predatory journals – typically pay-to-publish schemes with vaguely professional sounding names (American Journal of Science and Nature, etc), often ‘very rapid’ or no peer review, no editor, editorial board, or relationship with any institutions.

**Steps**

ETL project

1. ~~Generate a corpus of high quality biomedical research from pubmed~~

* ~~utilize pubmed API to collect data (~~[~~https://www.ncbi.nlm.nih.gov/books/NBK25497/~~](https://www.ncbi.nlm.nih.gov/books/NBK25497/)~~)~~
* ~~Search terms – COVID19, SARS-CoV-2, H1N1? Influenza? Immunology? Virology?~~

1. Using CORD-19, a pre-scraped dataset from Kaggle.

2. Generate a corpus of low quality biomedical research, will probably require more legwork (<https://predatoryjournals.com/journals/>).

- alternatively, use the journals listed here to see if any predatory journals are included in the CORD19 dataset

4. EDA on high and low quality articles

- length, frequency of new research, word clouds?, country of origin?

5. Automate scraping of medRxiv and BioRxiv preprints

ML Project

1. source a word-vector encoding that has been pre-trained on a large body of scientific literature. Even more ideally, this work would be trained specifically in biomedical research.

2. Settle on a methodology (transformer? LSTM? BoW?) and validate using

3. Implement methodology and test on known, example data (for instance, train on biotech vs other AP data and compare results)

Next steps

* I expect that there are many non-nlp based indicators of whether or not a given paper is likely to be high quality. For instance, authors who publish frequently in solid journals might be given some reputational credit. Conversely papers with very few references (or references which are mostly for the authors themselves) might be considered less trustworthy. Factoring these things in is basically a feature engineering projects that could improve the overall accuracy and might be fun to do.

Metrics:

Compare accuracy of differentiating high and low quality ALREADY PUBLISHED work from held out training set

Spot check some preprints?

Potential problems (would be fortunate to get this far)

Western bias

English bias

Innovation bias

**Scope for EDA**

One of the most amazing things about the response to the COVID-19 pandemic has been the truly incredible amount of research released on extremely short timelines. While this is an incredible human accomplishment, the generation of this much science exacerbates an existing problem facing many scientific researchers: it is often impossible to read every piece of potentially relevant research while still maintaining a productive output in the lab.

This problem is made even worse by the existence of relatively low quality science (whether intentionally fraudulent or simply careless) which must be waded through in order to find useful work. Ideally, it is the role of peer reviewed journals to solve this problem (though this process is [not perfect] (<https://statmodeling.stat.columbia.edu/2020/06/15/surgisphere-scandal-legacy-media-lancet-still-dont-get-it/> ) and journals are not without other flaws). However, simply publishing something in a professional journal is no guarantee that it is not ‘garbage science’. In fact, this whole project of evaluating the COVID-19 literature was inspired for me by

[this graphic](https://www.economist.com/graphic-detail/2020/05/30/how-to-spot-dodgy-academic-journals) from the economist, which shows that ~40% of the journals launched in 2018 were essentially little more than pay to publish scams. And this number has been growing since 2010! Clearly there is a market for writing and publishing junk science, and I see no reason why coronavirus related research would be exempt from this trend.