Predicting the Quality of COVID-19 Papers

Yuval Rubinstein, Prashanth Ramakrishna, Yi Yang, Aishwarya Manojkumar,
Emily Liang, Yash Bharti, Michelle Han

Introduction:

Hypothesis: It is possible to predict the quality or influence of a newly published research paper on COVID19.

Covid-19 was first identified in December 2019 as a cluster of pneumonia of unknown origin in Wuhan, China. It was then declared as a global pandemic in March 2020 by the World Health Organization. Since then, measures have been made to decrease the spread with mask protocol, social distancing, and overall health precautions. With over 71 million cases and 1.6 million deaths worldwide, the virus targets the older population, as well as those with pre-existing health conditions (such as cancer, chronic kidney disease, chronic obstructive pulmonary disease, or heart conditions). As a result, great strides in research regarding the virus have been made, creating a sudden influx of academic research papers. The urgency of this research program has resulted in two key problems. Firstly, this urgency has caused the quality of papers to suffer, not only because the research process itself is being rushed, but also because the quality-assurance process of peer-review is as well. For example, some articles' average time from submission to publication halved to about 60 days (according to Serge Horbach of Radboud University). Secondly, the high frequency of publication is posing problems for a research community that can only consume information so quickly. The ability, therefore, to determine which papers, upon publication, are of high quality and which are of low quality would be of high utility to the COVID19 research community. This would allow them to only pay attention to a small subset of high-quality papers, relieving the bottleneck of human attention, and simultaneously filter out, with high probability, low-quality work.

The goal of our research project is to determine whether predicting the quality of a newly published COVID19 research paper is possible. In the course of our investigation, we determined that while it may be possible to predict the quality or influence of a newly published paper, there are a number of practical limitations that make building a minimal viable product capable of such prediction currently intractable. In this report, we describe these limitations and possible ways to pivot in order to mitigate them.

Literature Review:

The notion of "quality" as it relates to scientific papers is difficult to define, particularly in the context of automated judgment. Conventionally, the quality of a paper is thought of as proportional to the impact it has had on the research community, quantified through citations. Our hypothesis can, therefore, to a certain degree, be reformulated as "It is possible to determine the long-term citation impact of newly published COVID19 research papers."

Determining the impact potential of research papers has been a highly studied topic. Research institutions, peer-reviewed journals, and indeed private companies are highly incentivized to be able to quickly differentiate the wheat from the chaff in relevant research domains. In recent years, due to the advent of machine learning and large digitally available troves of scientific research papers, work on this problem has indeed accelerated. Nevertheless, though progress has been made in hyper-specific subdomains, the problem, in its most general form, has remained unsolved.

A number of approaches to paper quality prediction are available. After surveying the relevant literature, we were ultimately left with a decision between two, detailed, respectively, in the following papers: "Predicting Rank for Scientific Research papers using supervised learning" (Mohamed El Mohadab, Belaid Bouikhalene, Said Safi)and "Predicting the long-term citation impact of recent publications" (Clara Stegehuis, Nelly Litvak, Ludo Waltman).

The goal of the research is to come up with a way to rank the quality of the scientific papers. To predict the rank of a newly published paper, this paper has created three different models using three differential learning methods: The paper chooses to work with the Multilayer Perceptron Algorithm, SMO Classifier, and Kstar Classifier. These classifiers are used as they allow for an easier analysis of the output data. The research results in a calculation of a rank based on a formula of multiple parameters such as author score, papers

published by authors and certain keywords. However, we decided not to go ahead with the modeling method in these publications due to the biases we learnt from the second research paper.

In the second paper that we analyzed (Predicting the long-term citation impact of recent publications"), we learnt of the multiple biases that exist in the current prediction methods. Variables such as the number of pages and the references in a paper can be easily manipulated. This allows researchers to artificially inflate these numbers. Some variables can be easily manipulated, such as the number of pages or references to other papers, this can lead to researchers artificially inflating certain parameters. This paper specifically analyzes the challenges with predicting long-term citation impact. It points out that using recent publications to predict the long-term citation impact has the same problem as those "sleeping-beauty" publications. "Sleeping-beauty" publications are those that are hardly cited for a long time and then suddenly receive a lot of citations. These kinds of publications make it very difficult to make accurate predictions. Another type of bias is the self-reinforcing effect. Self-reinforcing effect occurs due to preferential treatment given to the author's publication. So if an author was successful in his/her previous published works, then it will lead to a higher chance of success in the current work. As an example we can take a look at the previous variables which have been mentioned. For the 'Author Score' variable, self-reinforcement bias in an author's past successes will directly affect the rank of the paper. Even the 'Keywords & Average Publication/Keyword' variable can be easily manipulated by authors prior to publication.

Due to these biases, the goal of the research is to come up with an accurate probability distribution to predict the long-term citation impact (1-2 Years). The prediction mainly depends upon impact factors and citation impact. The probability distribution of the citation impact is calculated using quantile regression. Quantile regression is a type of regression analysis which estimates the conditional median of the response variable. Quantile regression is useful as its analysis is more robust against outlier measurements. Three models (probability distributions created using quantile regression) are evaluated in the paper, two models focus on creating a distribution based on impact factor and citations separately, then the last model focuses on predicting the distribution taking into account both the variables. The tails of these probability distributions are estimated by pareto tails. Pareto tails (a probability distribution which embodies the pareto principle of 80/20 rule) was expected as the majority of the citations belong to a few papers. After the creation of the models, their fit is compared. It is found that the model which uses both impact factor and citations is more accurate than the models which use only one parameter.

After viewing the results of this paper, we wanted to improve upon the probability distribution of the citation impact by introducing more features. Features such as degree centrality were added to create a more robust model. The decision to add degree centrality is inspired by other studies. These studies found conclusive results on a strong relationship between author's centrality in a co-authorship network and the citations received. The authors which have a higher centrality score generally received higher citations. We also added the number of downloads of a publication as a parameter, as well as number of readers according to a service such as Mendeley, and other types of altmetric indicators. Further research may investigate the effect of adding these predictors to our model. In particular, it would be interesting to find out whether the use of additional predictors decreases the level of uncertainty in predictions of long-term citation impact.

Methodology: Falsifiable Experiment and Data Collection:

In order for our featureset to have the most relevant information as well as be able to extract the information we need, we have come up with several criteria for which features are to be included. The features that were chosen all have common characteristics, such that they are difficult to manipulate, not self-reinforcing, reasonable quality proxies. We hoped that they would also be publicly available and easily traceable. Based on these criteria, four features were chosen: collaboration centrality score, number of citations, number of downloads, and journal of publication impact score. As suggested in related literature, some studies have been done in order to come up with a mathematical model for a ranking system, using several features that we ultimately decided against, including paper length, number of coauthors, number of keywords, and past performance indicators. As previously mentioned, these do not fit with our criteria of a good featureset. The biggest problem that was encountered during the features finding process was that in order to come up with the prediction, some metrics such as number of downloads and number of citations need to have information

available from the early days, as in soon after a journal is first published. That information was required for us to come up with a probability distribution of its citation impact based on quantile regression. However, the early data could not be found, which ultimately led to various problems, that will be discussed later down the line.

When Covid started spreading rapidly, the CDC began compiling research articles and journals into a database to help researchers find scholarly articles about COVID-19. The WHO started their own database and thereafter the CDC merged their database to the WHO's. Both of these databases had similar problems in terms of sourcing as well as formatting. Author names were formatted differently in different entries, sometimes comma delimited, other times semicolon delimited making it difficult to clean the dataset. The sourcing of both of these databases was based on keywords, so some entries in the dataset were not related to Covid at all. Due to these problems we decided to utilize the CORD-19 database. This database was created by the Semantic Scholar team at the Allen Institute for AI and had a total of 127k entries. The formatting in this database was uniform, making it easy to pull relevant information, however we still ran into the same sourcing problem. To mitigate this we used a SQL Query to filter the dataset by date. The date range we chose was March 1, 2020 to November 30, 2020. After that we generated a summary of the twenty journals that contributed the most to the database and ran another SQL query to generate a new set that only included articles from the top twenty journals that were published in the time frame specified. This brought our final dataset down to 10k entries. We wanted to limit the journals to those which contributed the most because there were many obscure journals that did not record metrics. Therefore to make our data pipeline more reliable we restricted the journal set and our data was no longer a random sampling. Doing this might have had unintended consequences so our data analysis later on tries to see whether this journal sampling produced biases in centrality score distribution.

Problems and Pivots:

Throughout our exploration of discovering the influence of a newly published COVID-19 paper, we faced many challenges that made it difficult for us to fully prove our hypothesis. The major areas that halted our progress were the inability to find earlier data on COVID-19 papers, the interdisciplinary nature of COVID, and the fact that the research on this topic is still in its infantile stages. Even though we were not able to tackle all of these issues during the short duration of our research project, we are optimistic that with more time, it is possible to create an analytic that expands on our research. The extra time will also allow more papers and more data points to come in that can be investigated. This approach in the future will be the pivot point of predicting the quality of COVID-19 papers.

During our project, we tried to amend the problems with the time granularity of data by using Web of Science, but ultimately found that resource to be flawed as well. While we were setting up our data pipeline, we found that we were unable to get data on paper metrics from previous time frames. Usually prediction takes data from the first year after a publication to predict the citation impact in 10 to 20 years in the future. COVID-19 and the research that corresponds have only been around for about 9 months. This short time range made it extremely difficult for us to perform our originally planned quantile regression on time series data of COVID-19 papers. We would have to rescale our time frame to fit the proportions of the short time that COVID-19 papers have been around. We explored a large variety of APIs for a number of databases (Elsevier API, Mendeley API, Altmetric API, CORD-19, PubMed, Google Scholar, SemanticScholar, etc.), but found that they all had different formats to their data. Some databases did not include the download numbers by paper, this was a feature that we needed for our model. Some included monthly citation data while others only had current citation data. All of these differences created massive inconsistencies when retrieving the metadata and rendered this portion of our data pipeline fruitless. However, a workaround for this particular obstacle could be the creation of a script that pulls metrics on a daily basis. We are currently working on this and this would not only create a more consistent data set in general, but also create a standard for future data granularity.

Another hurdle we must overcome before being able to fully tackle this hypothesis, is dealing with the discipline diversity of the studies of COVID-19. It is no surprise that COVID-19 has papers from journals in a broad range of disciplines. While it is beneficial for many areas to try to learn more about COVID-19, the outcome is that high quality journals and papers from disciplines that are not directly related to biology/medicine like physics may have an h-index that is lower than a low quality paper in a medical science journal. Physicists have tried to use a Gauss model to predict the projection of COVID-19, but this paper would

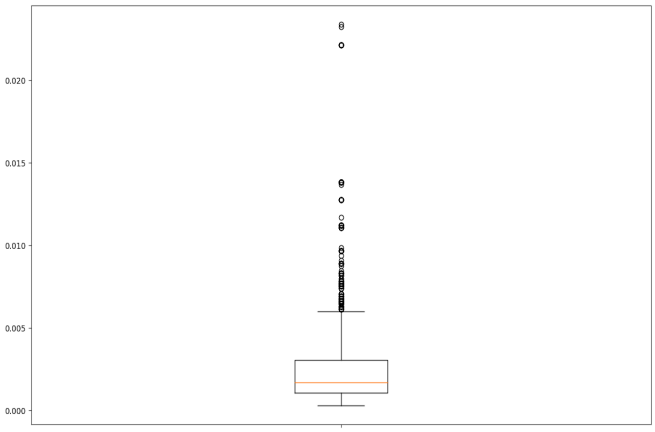
have a lower h-index because it is not where one would look first for this topic. This creates an inflation in healthcare related papers. Our automated h-index grab, worked on primarily by Aishwarya, found this discrepancy during the production of our data pipeline. This is a set back for our prediction model as we wanted to use the h-index of journals as a vital feature in our project. A solution to this paradox of h-indices that could be done immediately is designing a more precise dataset selection and pruning process. We did this by restricting the subset of journals that we looked at, but this creates a non-random sample that has bias and a skewed h-index distribution. If we were to use the h-indices from this subset as the feature to train our machine learning model, we would be unable to scale this and apply it to the larger dataset of other COVID-19 papers. Another way to go about this would be to divide journals by field of study and train our regression model that way. We can also separate the predictions by discipline and this would also create a more accurate prediction of quality for each individual batch of journals.

Possibly the most serious complication with this research topic is the freshness of this recently unstudied topic. Other domains of research like astrophysics and economics have had years to build a foundation of facts and theorems that are widely accepted. This allows us to have a base to compare new research to and be able to evaluate the paper's robustness. However, we do not have any of these established conditions for COVID-19. Without knowing the general consensus on the details of COVID-19, we have no guarantee that the current stream of research is reliable even if it has a high h-index or attains a lot of attention. This makes prediction of the long-term impact of a paper from the relatively short-term recognition not only hard to produce, but also hard to believe. There does not seem to be a quick-fix to this problem except to have patience and faith that scientists in the future will be able to work with a stronger foundation and build a dependable prediction model.

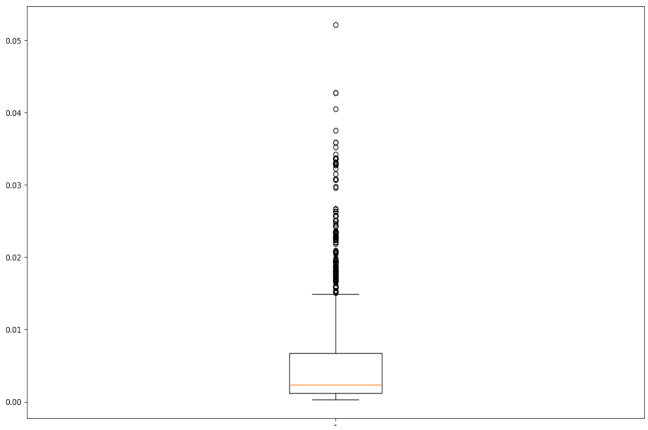
After working through the previously mentioned issues, we came to the realization that this research topic still needed more time to become more established. Even though the solution to obstacles we face is to essentially let time pass, we cannot be passive as any analysis and attempts are valuable to paving the road to the eventual success. Making a script that sets a high-standard precedent for granularity and metadata format, and making a model that trains and predicts off of journals separated by discipline are the crucial pivots to ultimately creating a well-working and respectable prediction model.

Data Analysis:

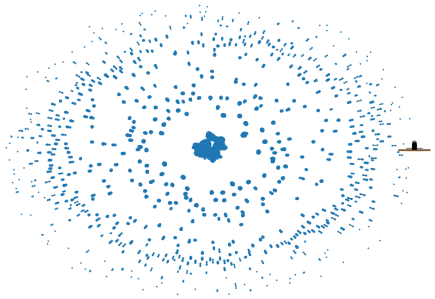
We did our data analysis comparing degree centrality and coauthorship network between the Journal subset and the full dataset. Since we used a non random sampling of papers due to ease of use for the data pipeline we plotted distributions of different features to see whether the subsetted data could be representative of the full dataset. There would be a flaw in our methodology if we trained the regression on a non representative data set because it would not reflect the bigger picture. We created a box and whisker plot to compare the degree centrality score distributions for the two datasets. Degree centrality is defined as the number of links incident upon a node (i.e., the number of ties that a node has). After analyzing the plots, the median centrality score of the journal subset is higher than the median score of the full dataset, meaning that authors were more connected in the journal subset than in the full dataset. Our second graph is the coauthor network for the Journal subset and the final graph is the coauthor network for the full dataset.



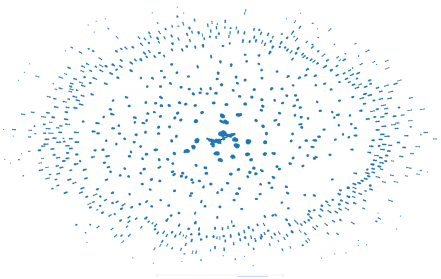
Degree Centrality of Full Dataset



Degree Centrality of Journal Subset



Coauthor Network of Journal Subset



Coauthor Network of Full Dataset

Collaboration Evaluation:

Prashanth came up with the project proposal. At the outset of the project, he produced a starting point from which the group could launch into a productive investigation. This included a team collaboration document where tasks and research progress could be kept track of, a literature review of relevant collected papers, a github repository for a common code base, and a script to generate collaboration networks from metadata. Over the course of the project, group members were fairly assigned work according to their ability and interest. Weekly meetings were set up in order to discuss progress on these tasks. However, almost no one completed the work assigned to them. As a result, in order for progress on the project to continue, Prashanth had to complete them. The group failed to collaborate well with each other. Attendance to weekly meetings was unreliable and group members failed to do even the basic work of reading the relevant papers to understand the problems at hand. These problems came to a head before the group presentation. Group members, having not contributed, were not able to create slides or confidently present. As a result, Prashanth created the majority of the presentation slides.

The project work distribution was skewed overwhelmingly towards Prashanth. The other group members that contributed tangible work products were Aishwarya and Yi. In particular, Aishwarya helped to generate plots exploring biases resulting from our non-random journal sampling, and Yi facilitated relations between group members. In order to make up for the skewed work distribution of the project and presentation slides, most of this report is written by all the group members outside of Prashanth. Michelle contributed the introduction, Yash and Ruby the literature review, Aishwarya MVP and Data Analysis, Emily the Problems and Pivots, and Yi this section. Prashanth outlined and edited the document.

Our inability to pivot over the course of the project was a direct consequence of poor collaboration. Failing fast and cheaply requires the ability to make quick progress towards specific questions. This would have

been possible if the work of a seven-person team were contributing to such quick progress. However, because only one person was making progress, we were unable to fail nearly as fast and cheaply as would have been necessary to productively pivot. This being said, we understand the importance of the hypothesis we hoped to investigate, and do believe that many observations we made over the course of our investigation are incredibly important, shedding light on difficult but potentially solvable limitations in predicting the long-term citation impact of recently published COVID19 research papers.

References

Abramo, Giovanni, et al. “Citations versus Journal Impact Factor as Proxy of Quality: Could the Latter Ever Be Preferable?” *Scientometrics*, vol. 84, no. 3, 27 Feb. 2010, pp. 821–833., doi:10.1007/s11192-010-0200-1.

Bai, Xiaomei et al. “An Overview on Evaluating and Predicting Scholarly Article Impact.” *Information* 8.3 (2017): 73. *Crossref*. Web.

Bento, Carolina, et al. “Predicting the Future Impact of Academic Publications.” *Progress in Artificial Intelligence Lecture Notes in Computer Science*, 2013, pp. 366–377., doi:10.1007/978-3-642-40669-0_32.

Hu, Xiaoli, et al. “Of Stars and Galaxies – Co-Authorship Network and Research.” *China Journal of Accounting Research*, Elsevier B.V., 29 Nov. 2019, www.sciencedirect.com/science/article/pii/S1755309119300358.

Mohadab, Mohamed El, et al. “Predicting Rank for Scientific Research Papers Using Supervised Learning.” *Applied Computing and Informatics*, Elsevier B.V., 6 Mar. 2018, www.sciencedirect.com/science/article/pii/S2210832717302703.

Noorden, Richard Van. “Formula Predicts Research Papers' Future Citations.” *Nature*, 3 Oct. 2013, doi:10.1038/nature.2013.13881.

Schüttler, J.; Schlickeiser, R.; Schlickeiser, F.; Kröger, M. Covid-19 Predictions Using a Gauss Model, Based on Data from April 2. Physics 2020, 2, 197-212.

Stegehuis, Clara, et al. “Predicting the Long-Term Citation Impact of Recent Publications.” *Journal of Informetrics*, vol. 9, no. 3, 31 Mar. 2015, pp. 642–657., doi:10.1016/j.joi.2015.06.005.