Predicting the Quality of Covid-19 Papers

Prashanth Ramakrishna (CEO & MVP), Yash Bharti, Yi Yang, Aishwarya Manojkumar, Emily Liang, Michelle Han, Yuval Rubinstein

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

Related Literature

Predicting the long-term citation impact of recent publications

- Quantile regression for predicting long term impact of papers from short term attention
- Prediction based on number of citations and journal of publication impact factor
- Prediction done within specific subject subdomains

Predicting rank for scientific research papers using supervised learning

- Learning Methods: Neural Networks, Hidden Markov Models, Support Vector Machines), Unsupervised (K-Means Clustering, Fuzzy C-Means
- Rank Equation:

$$Rank = \sum_{i=1}^{n} A_i . DL\left(\frac{1}{NH}\right) + 0.2(A_p + A_{nbr}) + 0.3(\Delta\omega + \text{type}) \pm \text{PR(i)} \times \text{DL}\left(\frac{1}{1 + \log A}\right)$$

Related Literature Cont.

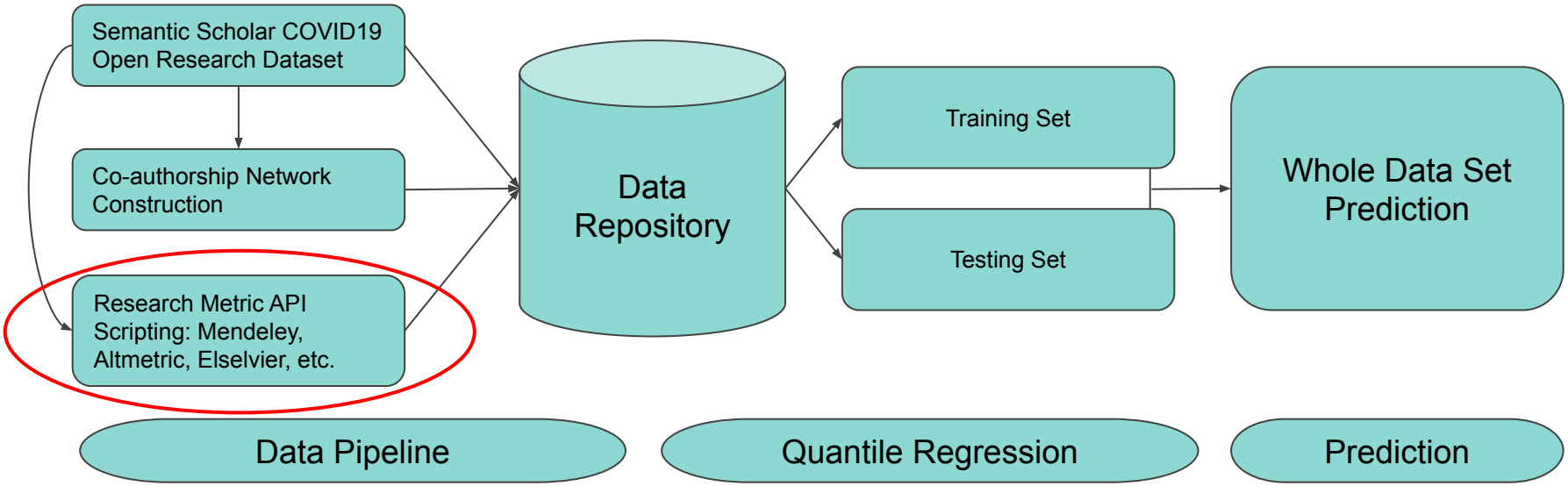
Citations versus journal impact factor as proxy of quality: could the latter ever be preferable?

- Understanding pros and cons of impact factor as quality proxy
- Calculation of “Scientific Strength” using both citations and impact factor
- Reliability of bibliometric proxy depending on “maturity” of associated information

Predicting scientific success based on coauthorship networks

- Social cognition and information filtering to study influence of networks on citation behavior
- Hindcasting: Assessing predictive power of author’s position in coauthorship network
- Supervised classification method based on Random Forest Classifier

Research Methodology



Feature Selection

Criteria

- Hard to manipulate
- Not self-reinforcing
- Reasonable quality proxy
- Easily track-able over time
- Publicly available

TIME FRAME: Data for features 1-3 is required for “early days”, defined via scaling

Included Features

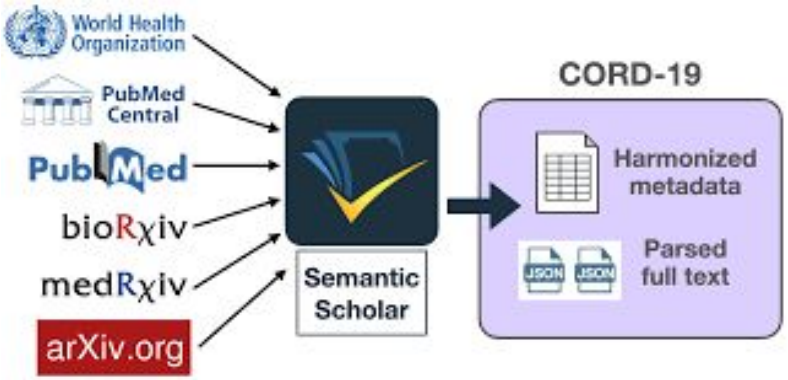
1. Collaboration Centrality Score
2. Number of Downloads
3. Number of Citations
4. Journal of Pub. Impact Score

Avoided Features

- Paper Length
- Number of Coauthors
- Key Words
- Past Performance Indicators

Dataset Selection and Pruning

- CDC vs WHO vs CORD-19
- SQL Queries for publication date
- Restriction to targeted journals
- ~127K → ~10K

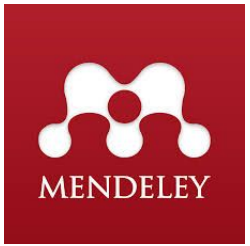


CORD-19

COVID-19 Open Research Dataset

The Semantic Scholar team at the Allen Institute for AI has partnered with leading research groups to provide CORD-19, a free resource of more than 280,000 scholarly articles about the novel coronavirus for use by the global research community.

Data Pipeline Congestion



Obstacles

- APIs only had current metrics
- APIs had incomplete metrics
- Not all papers had available metrics
- Paper granularity inconsistent between metrics
- Much of metadata kept by year/month

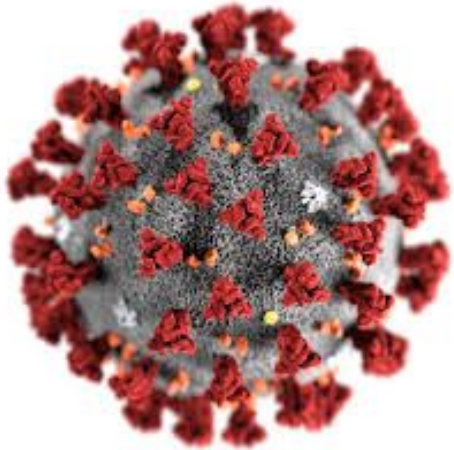
Data Pipeline Unclogging

Problems	Solutions
Could not access data on paper metrics from previous timeframes	Using paid Web of Science subscription for our dataset
Current metrics available = unable to predict a paper's citation impact in the long term	Script to track metrics daily → time continuous feature data (~5 months for trainable dataset)
Paper's have not been published for a long enough time	Rescaling the time frame of the data set appropriately

Uncertainty at Birth

The World Health Organization temporarily halted clinical trials using hydroxychloroquine after a Lancet article was originally published, a move that reflected the influence a single study can have in the fast-changing area of coronavirus research

- During the earlier stages, little was known about the virus
- Newer papers are better by default due to nascent knowledge acquisition: new information comes out, leading to better quality papers that better grounding
- A training set consisting of earlier papers will have a hard time predicting quality of later papers, since a domain foundation has not yet stabilized



Data Analysis

Coauthorship Network Exploration

title STRING	publish_time DATE	authors STRING	journal STRING
Ethical decision making in a pandemic: where are the voices of vulnerable people?	2020-06-19	McCullough, Melissa	BMJ
CQC says inspections suspended for covid-19 crisis will restart in autumn.	2020-06-19	Iacobucci, Gareth	BMJ
UK's response to covid-19: crude, unadjusted mortality figures are not the whole story.	2020-06-19	Greenberg, Aryeh L; Greenberg, Harry	BMJ
Covid-19: UK drops its own contact tracing app to switch to Apple and Google model.	2020-06-19	Wise, Jacqui	BMJ
Covid-19: now is not the time to judge the UK's response.	2020-06-19	Slingo, Mary	BMJ

Raw Dataset Exploration

Conclusion: Hypothesis False … For Now

References

1. 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.
2. Noorden, Richard Van. “Formula Predicts Research Papers' Future Citations.” *Nature*, 3 Oct. 2013, doi:10.1038/nature.2013.13881.
3. 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.
4. 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.
5. 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.
6. Bai, Xiaomei et al. “An Overview on Evaluating and Predicting Scholarly Article Impact.” *Information* 8.3 (2017): 73. *Crossref*. Web.
7. 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.