A Synthetic Prediction Market for Estimating Confidence in Published Work

Sarah Rajtmajer,[1] Christopher Griffin,[1] Jian Wu,[2] Robert Fraleigh,[1] Laxmaan Balaji,[1] Anna Squicciarini,[1] Anthony Kwasnica,[1] David Pennock,[3] Michael McLaughlin,[1] Timothy Fritton,[1] Nishanth Nakshatri,[1] Arjun Menon,[1] Sai Ajay Modukuri,[1] Rajal Nivargi,[1] Xin Wei,[2] C. Lee Giles[1]

[1]The Pennsylvania State University
[2]Old Dominion University
[3]Rutgers University
{smr48,cxg286,rdf5090,lpb5347,acs20,amk17,mvm7085,tjf115,nzn5185,amm8987,svm6277,rfn5089,clg20}@psu.edu
{j1wu,xwei001}@odu.edu
david.pennock@rutgers.edu

Abstract

Explainably estimating confidence in published scholarly work offers opportunity for faster and more robust scientific progress. We develop a synthetic prediction market to assess the credibility of published claims in the social and behavioral sciences literature. We demonstrate our system and detail our findings using a collection of known replication projects. We suggest that this work lays the foundation for a research agenda that creatively uses AI for peer review.

Introduction

Concerns about the replicability, robustness and reproducibility of findings in scientific literature have gained widespread attention over the last decade in the social sciences and beyond, including AI (Gundersen and Kjensmo 2018; Henderson et al. 2018; Hutson 2018; Haibe-Kains et al. 2020; Pineau et al. 2021). This attention has been catalyzed by and has likewise motivated a number of large-scale replication projects (Open Science Collaboration 2015; Camerer et al. 2016, 2018; Klein et al. 2014, 2018; Cova et al. 2021) which have reported successful replication rates anywhere between 36% and 78% and have further escalated debate of a crisis of confidence in present-day empirical work (Baker 2016; Gilbert et al. 2016; Fanelli 2018).

Given the challenges and significant resources required to run high-powered replication studies, researchers have sought other approaches to assess confidence in published claims and have looked to creative assembly of expert judgement as one opportunity. Initial evidence has supported the promise of prediction markets in this context (Dreber et al. 2015; Camerer et al. 2016, 2018; Forsell et al. 2019; Gordon et al. 2020, 2021). However, practical deployment of prediction markets to evaluate scientific findings is also limited. They require the coordinated, sustained effort of collections of human experts. They typically rely on availability of some measurement of ground truth. That is, participants trade on well-defined, verifiable outcomes determined after market close (although, see (Liu, Wang, and Chen 2020) for recent work proposing a surrogate scoring mechanism).

Copyright © 2022, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Another set of limitations centers around the shortcomings of human market participants. Researchers base their assessments on the work with which they are familiar, the reputations of journals, and similar. Their judgements may be influenced by cognitive biases, e.g., anchoring, confirmation bias (Fraser et al. 2021), and the compounded effects of these biases in market settings are poorly understood.

We suggest that markets populated by artificial agents provide an opportunity to overcome or mitigate many of these limitations. Synthetic prediction markets can be deployed rapidly and at scale. Artificial agents can have broad access to the literature and metadata at scales far beyond the capacity of an individual researcher.

The system we demonstrate here is a fully synthetic prediction market wherein algorithmic agents (trader bots) are trained and tested on proxy ground truth pulled from existing replication studies. Our work is complementary to recent efforts using machine learning for reproducibility prediction (Altmejd et al. 2019; Yang, Youyou, and Uzzi 2020; Pawel and Held 2020; Wu et al. 2021). Unlike prior approaches the market scores only a subset of the papers in our test set, but accuracy on that subset is very high. The market-based approach affords explainability by way of the record of trades and corresponding relevant features.

System

The prototype system is built around two primary modules, namely, a feature extraction pipeline and the synthetic market. Outputs of feature extraction are provided to agents which populate the market during train and test (Figure 1).

Feature Extraction Pipeline

The Feature EXtraction framework for Replicability prediction (FEXRep) extracts five categories of features related to a given scholarly preprint or published paper and its metadata: biblometric, venue-related, author-related, statistical and semantic information. At present, 41 total features are extracted, ranging from p values and sample size to number of authors and acknowledgement of funding. Further detail is provided in (Wu et al. 2021).

In the prototype system, all features represent paper-level information. Ongoing efforts are expanding extraction to in-

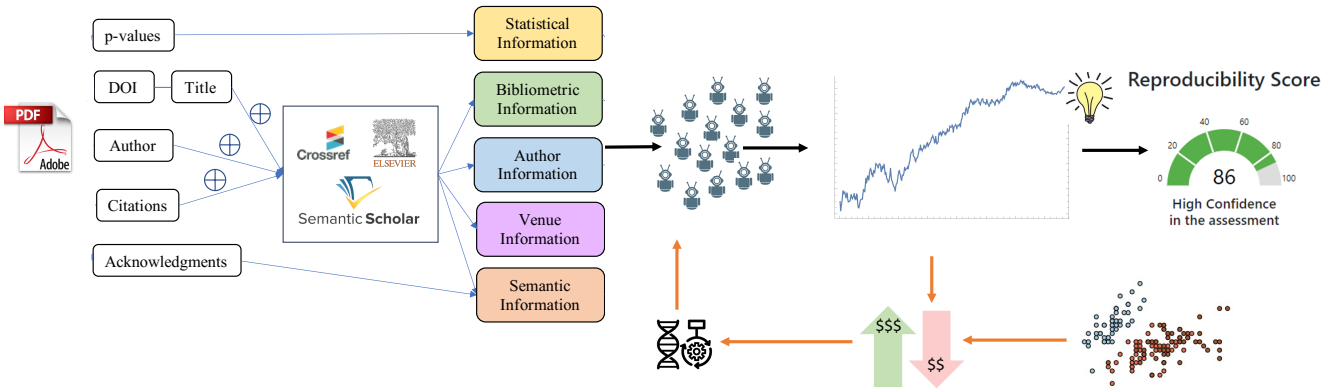

Figure 1: (Black arrows) A scientific paper is processed through the FEXRep feature extraction framework. Features are shared with the agents who purchase assets corresponding to binary outcomes of a notional replication study of the primary claim of that paper. The price of these assets at market close is an indicator of confidence in the claim. (Orange arrows) During training, agents purchase assets corresponding to claims drawn from prior replication projects for which ground truth is known. At the close of each market, some agents profit and others lose money. An evolutionary algorithm is used to update the population.

corporate features at the claim-level. This will allow for individual assessment of multiple claims within the same paper, rather than the current approach which scores the primary claim of the paper as it is asserted in the abstract.

Synthetic Market

Agents in the market are initialized with a fixed amount of cash and provided with the set of extracted features representing a paper in question. Agents may purchase assets corresponding to *will replicate* or *will not replicate* outcomes of a notional replication study of the primary claim of that paper. Agent purchase logic is defined using a sigmoid transformation of a convex semi-algebraic set defined in feature space. Asset prices are determined by a logarithmic scoring rule, and for simplicity, agents specialize in purchasing one of the two asset classes. Time-varying asset prices affect the structure of the semi-algebraic sets leading to time-varying agent purchase rules (see (Nakshatri et al. 2021) for further detail including theoretical properties of the market). The price of a *will replicate* asset at market close is taken as proxy for confidence in the primary claim of the paper.

During training, parameters that define agent purchase logic are identified using an evolutionary algorithm. The objective function minimizes root mean square error of the estimated score. Agent performance is evaluated by profit made. Profitable agents are retained, allowed to replicate and then modified using mutation and crossover of parameter values. Agents that do not make a profit are deleted.

Explainability

The current prototype provides explanations of scores through the record of agents participating and trades made. Confidence in the system's assessment of a paper is based on the extent of agent participation. Agents are initialized in different positions within feature space, so the trading patterns of each agent can be explained in terms of their position and the geometry defining their purchase logic.

Evaluation

Initial testing of our prototype system was done using a collection of known replication projects and outcomes. In particular, we use the Reproducibility Project Psychology (Open Science Collaboration 2015), Social Science Replication Project (Camerer et al. 2018), Experimental Economics Replication Project (Camerer et al. 2016), Many Labs (Klein et al. 2014) and Many Labs 2 (Klein et al. 2018). Collectively, those projects represent primary findings of 192 total papers in the social and behavioral sciences, each labeled either *Replicable* or *Not Replicable*.

Experimental settings. Five-fold cross validation was used. Each fold contained 153 training and 39 test points. Initial conditions were fixed over the five folds – specifically, we seeded 5 agents per market, each was given 5 units of cash, and the initial price of a *will replicate* asset was set to 0.5. The genetic algorithm trained over 50 generations.

Results on scored papers. Our system provides a confidence score for 68 of 192 (35%) of the papers in our set. On the set of scored papers, accuracy is 0.894, precision is 0.917, recall is 0.903, and **F1** is **0.903** (macro averages). A sizeable un-scored subset of data (65%) is the trade-off for high accuracy on the scored subset of the data. A test point is un-scored when the system has determined it has insufficient information to evaluate it.

System non-scoring. Unlike most other machine learning algorithms, the synthetic market does not provide an evaluation for every input. Like its human-populated counterparts, the market is vulnerable to lack of participation (Arrow et al. 2008; Tetlock 2008; Rothschild and Pennock 2014). Agents will not participate if they have not seen a sufficiently similar training point (paper). This is more common when the training dataset is small; in experiments with larger datasets, we have observed participation increases. Meaningful ways to increase agent participation, including hybrid settings with human participants, are being explored.

Acknowledgements

We acknowledge support by DARPA W911NF-19-2- 0272. This work does not necessarily reflect the position or policy of DARPA and no official endorsement should be inferred.

References

Altmejd, A.; Dreber, A.; Forsell, E.; Huber, J.; Imai, T.; Johannesson, M.; Kirchler, M.; Nave, G.; and Camerer, C. 2019. Predicting the replicability of social science lab experiments. *PloS one*, 14(12): e0225826.

Arrow, K. J.; Forsythe, R.; Gorham, M.; Hahn, R.; Hanson, R.; Ledyard, J. O.; Levmore, S.; Litan, R.; Milgrom, P.; Nelson, F. D.; Neumann, G. R.; Ottaviani, M.; Schelling, T. C.; Shiller, R. J.; Smith, V. L.; Snowberg, E.; Sunstein, C. R.; Tetlock, P. C.; Tetlock, P. E.; Varian, H. R.; Wolfers, J.; and Zitzewitz, E. 2008. The Promise of Prediction Markets. *Science*, 320(5878): 877–878.

Baker, M. 2016. 1,500 scientists lift the lid on reproducibility. *Nature News*, 533(7604): 452.

Camerer, C. F.; Dreber, A.; Forsell, E.; Ho, T.-H.; Huber, J.; Johannesson, M.; Kirchler, M.; Almenberg, J.; Altmejd, A.; Chan, T.; et al. 2016. Evaluating replicability of laboratory experiments in economics. *Science*, 351(6280): 1433–1436.

Camerer, C. F.; Dreber, A.; Holzmeister, F.; Ho, T.-H.; Huber, J.; Johannesson, M.; Kirchler, M.; Nave, G.; Nosek, B. A.; Pfeiffer, T.; et al. 2018. Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, 2(9): 637–644.

Cova, F.; Strickland, B.; Abatista, A.; Allard, A.; Andow, J.; Attie, M.; Beebe, J.; Berniūnas, R.; Boudesseul, J.; Colombo, M.; et al. 2021. Estimating the reproducibility of experimental philosophy. *Review of Philosophy and Psychology*, 12(1): 9–44.

Dreber, A.; Pfeiffer, T.; Almenberg, J.; Isaksson, S.; Wilson, B.; Chen, Y.; Nosek, B. A.; and Johannesson, M. 2015. Using prediction markets to estimate the reproducibility of scientific research. *Proceedings of the National Academy of Sciences*, 112(50): 15343–15347.

Fanelli, D. 2018. Opinion: Is science really facing a reproducibility crisis, and do we need it to? *Proceedings of the National Academy of Sciences*, 115(11): 2628–2631.

Forsell, E.; Viganola, D.; Pfeiffer, T.; Almenberg, J.; Wilson, B.; Chen, Y.; Nosek, B. A.; Johannesson, M.; and Dreber, A. 2019. Predicting replication outcomes in the Many Labs 2 study. *Journal of Economic Psychology*, 75: 102117.

Fraser, H.; Bush, M.; Wintle, B.; Mody, F.; Smith, E.; Hanea, A.; Gould, E.; Hemming, V.; Hamilton, D.; Rumpff, L.; et al. 2021. Predicting reliability through structured expert elicitation with repliCATS.

Gilbert, D. T.; King, G.; Pettigrew, S.; and Wilson, T. D. 2016. Comment on "Estimating the reproducibility of psychological science". *Science*, 351(6277): 1037–1037.

Gordon, M.; Viganola, D.; Bishop, M.; Chen, Y.; Dreber, A.; Goldfedder, B.; Holzmeister, F.; Johannesson, M.; Liu, Y.; Twardy, C.; et al. 2020. Are replication rates the same across academic fields? Community forecasts from the DARPA SCORE programme. *Royal Society open science*.

Gordon, M.; Viganola, D.; Dreber, A.; Johannesson, M.; and Pfeiffer, T. 2021. Predicting replicability—Analysis of survey and prediction market data from large-scale forecasting projects. *Plos one*, 16(4): e0248780.

Gundersen, O. E.; and Kjensmo, S. 2018. State of the art: Reproducibility in artificial intelligence. In *Thirty-second AAAI conference on artificial intelligence*.

Haibe-Kains, B.; Adam, G. A.; Hosny, A.; Khodakarami, F.; Waldron, L.; Wang, B.; McIntosh, C.; Goldenberg, A.; Kundaje, A.; Greene, C. S.; et al. 2020. Transparency and reproducibility in artificial intelligence. *Nature*, 586(7829).

Henderson, P.; Islam, R.; Bachman, P.; Pineau, J.; Precup, D.; and Meger, D. 2018. Deep reinforcement learning that matters. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Hutson, M. 2018. Artificial intelligence faces reproducibility crisis. *Science*, 359: 478.

Klein, R. A.; Ratliff, K. A.; Vianello, M.; Adams Jr, R. B.; Bahník, Š.; Bernstein, M. J.; Bocian, K.; Brandt, M. J.; Brooks, B.; Brumbaugh, C. C.; et al. 2014. Investigating variation in replicability. *Social psychology*.

Klein, R. A.; Vianello, M.; Hasselman, F.; Adams, B. G.; Adams Jr, R. B.; Alper, S.; Aveyard, M.; Axt, J. R.; Babalola, M. T.; Bahník, Š.; et al. 2018. Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, 1(4): 443–490.

Liu, Y.; Wang, J.; and Chen, Y. 2020. Surrogate scoring rules. In *Proceedings of the 21st ACM Conference on Economics and Computation*, 853–871.

Nakshatri, N.; Menon, A.; Giles, C. L.; Rajtmajer, S.; and Griffin, C. 2021. Design and Analysis of a Synthetic Prediction Market using Dynamic Convex Sets. *arXiv preprint arXiv:2101.01787*.

Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science*, 349(6251).

Pawel, S.; and Held, L. 2020. Probabilistic forecasting of replication studies. *PloS one*, 15(4): e0231416.

Pineau, J.; Vincent-Lamarre, P.; Sinha, K.; Larivière, V.; Beygelzimer, A.; d'Alché Buc, F.; Fox, E.; and Larochelle, H. 2021. Improving Reproducibility in Machine Learning Research. *Journal of Machine Learning Research*, 22.

Rothschild, D.; and Pennock, D. M. 2014. The extent of price misalignment in prediction markets. *Algorithmic Finance*, 3(1-2): 3–20.

Tetlock, P. C. 2008. Liquidity and prediction market efficiency. *Available at SSRN 929916*.

Wu, J.; Nivargi, R.; Lanka, S. S. T.; Menon, A. M.; Modukuri, S. A.; Nakshatri, N.; Wei, X.; Wang, Z.; Caverlee, J.; Rajtmajer, S. M.; et al. 2021. Predicting the Reproducibility of Social and Behavioral Science Papers Using Supervised Learning Models. *arXiv preprint arXiv:2104.04580*.

Yang, Y.; Youyou, W.; and Uzzi, B. 2020. Estimating the deep replicability of scientific findings using human and artificial intelligence. *Proceedings of the National Academy of Sciences*, 117(20): 10762–10768.