RES: An Interpretable Replicability Estimation System for Research Publications

Zhuoer Wang[1]*, Qizhang Feng[1]*, Mohinish Chatterjee[1]*, Xing Zhao[1], Yezi Liu[1], Yuening Li[1], Abhay Kumar Singh[1], Frank M. Shipman[1], Xia Hu[2], James Caverlee[1]

[1]Department of Computer Science and Engineering, Texas A&M University
[2]Department of Computer Science, Rice University
{wang, qf31, mohi_chat, xingzhao, yeziliu, liyuening, abhay, shipman, caverlee}@tamu.edu
xia.hu@rice.edu

Abstract

Reliable and faithful research is the cornerstone of breakthrough advancements and disruptive innovations. Assessing the credibility of scientific findings and claims in research publications has long been a time-consuming and challenging task for researchers and decision-makers. In this paper, we introduce RES - an intelligent system that assists humans in analyzing the credibility of scientific findings and claims in research publications in the field of social and behavioral sciences by estimating their replicability. The pipeline of RES consists of four major modules that perform feature extraction, replicability estimation, result explanation, and sentiment analysis respectively. Our evaluation based on human experts' assessments suggests that the RES has achieved adequate performance. The RES is also built with a Graphical User Interface (GUI) that is publicly accessible at https://tamu-infolab.github.io/RES/.

Introduction

Replicability is central to the evaluation of research credibility. Findings and claims made in unreplicable research may mislead subsequent researchers and decision-makers, resulting in enormous social and economic impacts. Due to the growing concerns of research replicability across many fields (Camerer et al. 2016; Aarts et al. 2015; Altmejd et al. 2019), efforts such as the Open Science Collaboration (Lakens et al. 2012) and Many Labs (Klein et al. 2014, 2018) have led pioneering efforts to directly replicate experiments from some high profile studies.

While promising, these efforts are expensive, time-consuming, and require significant lag times from initiation to final results. Hence, there is growing interest in relying on expert assessments of a study's potential replicability to provide rapid feedback. Several studies suggest that experienced human experts can conclude from the content of the original paper about which findings are likely to replicate, often relying on the findings and the associated supporting evidence (Dreber et al. 2015; Gordon et al. 2020; Fraser et al. 2021). Naturally, these human assessments when paired with advances in machine learning offer the tantalizing possibility of real-time inference of the replicability of published research (Forsell et al. 2019; Wu et al. 2021). For example, Altmejd et al. examined several black-box statistical models that make binary replication predictions based on features representing statistical experimental design and result properties, outcomes, citation counts, author metrics, and subjects. Yang, Youyou, and Uzzi further integrated narrative text into a neural model. Most recently, the DARPA Systematizing Confidence in Open Research and Evidence (SCORE) program (Alipourfard et al. 2021) launched an effort to build comprehensive models for the prediction of replicability through data collected from systematically conducted re-experiments and experts' annotations.

In this paper, we present RES, an interactive, intelligent, and publicly accessible system that provides real-time automated estimation of replicability. The RES makes three contributions: 1) It exploits a more exhaustive set of features and leverages a larger training set; 2) It supplies more transparency to the estimation model via explainable results; 3) It supports the sentiment analysis of mentions in subsequent publications, which offers valuable opinions from experts in the research fields.

The RES System

In this section, we detail the design and features of the RES. The RES is built with Bootstrap , which provides an easy-to-use interactive web-based user interface. The system takes a CSV file that contains the publication's title, digital object identifier, and testing claims as input. Then, the system executes each module and displays associated information upon the user's request, as illustrated in Figure 1.

Replicability Estimation

To estimate the replicability of research publications, we exploited an extensive set of intrinsic and extrinsic features associated with each publication. Intrinsic features represent the publication's content, experimental design, associated results, and scientific claims. Intrinsic features typically formalize a fundamental profile of the research that determines its replicability, and they can be extracted directly from the content of the research publication. In addition to the textual claim content input by the user, the RES also extracts content directly from the publication through FREX [1], an

*These authors contributed equally.

[1] https://github.com/amm-kun/score_psu

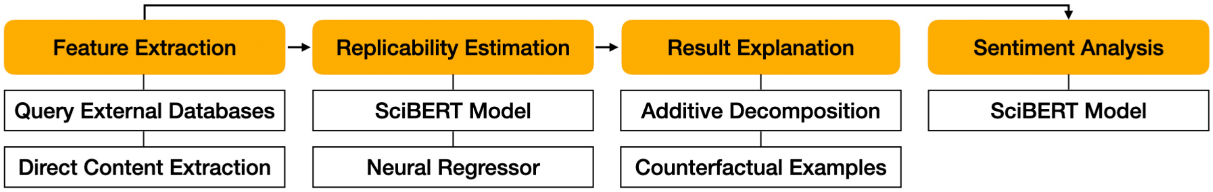

Figure 1: RES System Pipeline.

.

Type	Category	Source	Representing Features
Extrinsic	Bibliometrics	Semantic Scholar or Google Scholar	Citation Count, Citation Velocity, Influential Citation Count
	Author Profile		Total Publications, Total Citations, h-index
	Venue Profile	SCOPUS	Scholarly Output, Source Normalized Impact, Journal Ranking
Intrinsic	Experiment Statistics	FREX	Significance, p-value, Sample Size
	Textual Content	User Input	Claim Abstract, Hypothesis, Test Specifications

Table 1: Features used for the estimation of replicability

open-source information extraction package developed under the SCORE program. Extrinsic features are subjective facts that suggest the influence of the research publication and the academic community's evaluation of the research publication. Extrinsic features are usually indirect clues of the research's replicability, and they can be retrieved through querying databases like Semantic Scholar , Google Scholar , and SCOPUS[2]. We categorized intrinsic and extrinsic features into five different types and listed their source as well as representing features in Table 1. The RES provides a look-up table that contains detailed feature explanations through its GUI. After feature preparation, the RES feeds those data to the trained replicability estimation model. The model consists of a neural structured data regressor and a pretrained language model SciBERT (Beltagy, Lo, and Cohan 2019) . We trained the replicability estimation model with 2,400 samples containing the same set of features as described in Table 1 and annotations of replicability scores made available through the SCORE program (Fraser et al. 2021) as the ground truth label. The training objectives of the model are minimizing the mean squared error and maximizing the ranking correlation. The model achieved 0.137 RMSE and 0.32 Spearman ranking correlation on the test set.

Explainability

The explainability component aims at providing transparency to the RES. Two modules, the clause explanation module and counterfactual explanation module, collaborate together to offer explanation for textual and metadata features.

Attribution of recurrent neural network predictions via additive decomposition The clause level explanation module is built upon REAT (Du et al. 2019), which is a post-hoc explanation method. To explain the system decision at clause level, we take a text claim and a pre-trained model as input. The explanation result will be presented in the form of a heatmap, where the color represents the direction and the depth of the color indicates the contribution.

Understanding Black-Box Model Predictions by Counterfactual Explanation The counterfactual explanation module is built upon DiCE (Mothilal, Sharma, and Tan 2020), which explains the decision of ML-based systems via counterfactual examples. Using metadata features and a trained model as input, RES can effectively deliver various counterfactual examples to the users. In addition, the interactive GUI allows users to explore possible model decisions through customised meta-features. Moreover, RES provides contrastive explanation implemented based on (Anjomshoae, Främling, and Najjar 2019). Taking a sample and its counterfactual example, our method compares the scrambled features in terms of contextual importance and utility.

Downstream Sentiment Analysis

We implemented an auxiliary downstream sentiment analysis module to provide the user with additional insights for assessing the research publication's replicability. The module extracts and analyzes textual mentions of the targeting research publication in papers that have cited it. Specifically, We annotated 3,060 downstream mentions based on their replicability-oriented sentiments and trained a SciBERT-based model that can classify a downstream mention into one of the following three classes: Positive, Negative, and None sentiment. According to the evaluation on a holdout set, the model achieved 0.7419 Macro F1.

Conclusions and Future Work

We built an interactive, intelligent, and publicly accessible system that estimates replicability with more features, more samples, better explainability, and the capability of analyzing downstream mentions. In the future, we will adapt RES to domains beyond the field of social and behavioral sciences. We will also improve our models for better estimation performance and transparency.

Acknowledgements

The authors thank the funding agencies for their generous supports. The work is in part supported by DARPA grant W911NF19-2-0272. The views and conclusions in this paper are those of the authors and should not be interpreted as representing any funding agencies.

[2] https://www.scopus.com/

References

Aarts, A.; Anderson, J.; Anderson, C.; Attridge, P.; Attwood, A.; Axt, J.; Babel, M.; Bahník, Š.; Baranski, E.; Barnett-Cowan, M.; et al. 2015. Estimating the reproducibility of psychological science. *Science*, 349(6251).

Alipourfard, N.; Arendt, B.; Benjamin, D. M.; Benkler, N.; Bishop, M. M.; Burstein, M.; Bush, M.; Caverlee, J.; Chen, Y.; Clark, C.; and et al. 2021. Systematizing Confidence in Open Research and Evidence (SCORE).

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.

Anjomshoae, S.; Främling, K.; and Najjar, A. 2019. Explanations of black-box model predictions by contextual importance and utility. In *International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems*, 95–109. Springer.

Beltagy, I.; Lo, K.; and Cohan, A. 2019. SciBERT: A Pretrained Language Model for Scientific Text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3615–3620.

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.

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.

Du, M.; Liu, N.; Yang, F.; Ji, S.; and Hu, X. 2019. On attribution of recurrent neural network predictions via additive decomposition. In *The World Wide Web Conference*, 383–393.

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. T.; Hanea, A.; Gould, E.; Hemming, V.; Hamilton, D. G.; Rumpff, L.; and et al. 2021. Predicting reliability through structured expert elicitation with repliCATS (Collaborative Assessments for Trustworthy Science).

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*.

Klein, R.; Ratliff, K.; Vianello, M.; Adams Jr, R.; Bahnik, S.; Bernstein, M.; Bocian, K.; Brandt, M.; Brooks, B.; Brumbaugh, C. C.; et al. 2014. Investigating variation in replicability: A "Many Labs" replication project. *Social Psychology*, 45(3): 142–152.

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.

Lakens, D.; et al. 2012. An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science*, 7(6): 657–660.

Mothilal, R. K.; Sharma, A.; and Tan, C. 2020. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 607–617.

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.