Introduction to the HICSS-55 Collaboration for Data Science Minitrack

Lina Zhou
Business Information Systems and Operations Management
The University of North Carolina at Charlotte
Charlotte, NC 28223
lzhou8@uncc.edu

Souren Paul
Graduate School of Computer & Information Sciences
Nova Southeastern University
3301 College Avenue
Fort Lauderdale, FL 33314
Souren.paul@gmail.com

Florian Schwade
University of Koblenz-Landau
Faculty of Computer Science
Institute for IS Research
Universitaetsstrasse 1
D-56070 Koblenz
fschwade@uni-koblenz.de

Data science projects are complex in terms of the set of skills, knowledge, even experience that they require. Collaboration is a key factor in addressing the complexity of data science projects to improve the process and outcomes of decision making. According to a recent survey of 183 IBM employees who had experience in data science [1], they collaborate extensively with different roles on data science teams, including engineer/analyst/programmer, communicator, researcher/scientist, manager/executive, and domain experts. In addition, the collaboration runs through the entire workflow of a data science project, ranging from creating measurement plan, through preparing data, training and applying model(s), and evaluating outcomes, to communicating with stakeholders. In particular, some roles such as researcher/scientist, engineer/analyst/programmer, and domain experts, were consistently involved in collaboration across the different stages of a data science project [1]. Collaboration contributes to the productivity and efficiency of data science teams by identifying relevant questions or problems, collecting and integrating data from different sources, managing and making sense of the data by building models, evaluating the models based on selected measures, communicating findings in ways that are easily understandable, and applying the models to create real impacts. Furthermore, the roles engaged in the data science collaboration are not limited to team members but may also include other direct or indirect stakeholders and even AI systems.

There has been a continued interest in finding ways to increase the value of data science and use it to address business challenges and enhance their operational efficiency and/or competitive advantages. This minitrack includes one paper session, consisting of two papers covering the following areas of interest: collaborative data science, collaboration for social impact of data science, collaborative analysis of big data, social and psychological issues in collaborative data science, social media driven collaborative data science, collaborative crowdsourcing analytics, and human-machine collaboration in data science.

The first paper, "The Impacts of Lockdown on Open Source Software Contributions during the COVID-19 Outbreak", examines individuals' contributions to open source software on GitHub in the Wuhan region affected by the COVID-19 lockdown. The paper is based on a natural experiment using developers in the HMT region, which locked down much later, as a control group. Based on comparing the contribution levels of Wuhan and HMT developers before and after the lockdown, the paper studies the change in Wuhan developers contributions. The analysis of the GitHub commits and comments shows that Wuhan developers made fewer contributions than HMT developers in the first five weeks after the lockdown in Wuhan. The paper concludes that the reduced GitHub contributions in the Wuhan region may result from a lack of face to face collaboration during the lockdown.

The second paper, "Tune Down the Misinformation, Please: Generating Corrective Messages for COVID-19 Misinformation", proposes an analytical pipeline for semi-automatically generating corrective messages toward COVID-19 misinformation in social media. The proposed analytical pipeline encompasses ingestion of inference data, misinformation detection and misinformation intervention,ultimately resulting in candidate corrective messages. As the pipeline contains an additional detection enhancement module, that is trained based on a random sample of the inference data, the misinformation can be enhanced. The evaluation of this pipeline against a large data set confirms its efficiency. The paper makes the case that the analytical pipeline can be combined with human intelligence to combat misinformation concerning COVID-19 in social media.

URI: https://hdl.handle.net/10125/79367
978-0-9981331-5-7

References

[1] A. X. Zhang, M. Muller, and D. Wang, "How do Data Science Workers Collaborate? Roles, Workflows, and Tools," *Proc. ACM Hum.-Comput. Interact.,* vol. 4, no. CSCW1, p. Article 022, 2020.