Data Minding and Mining: Summary of Methods and Significance

This paper builds on *Enhancing (Publications on) Data Quality: Deeper Data Minding and Fuller Data Confession*, published by Xiao-Li Meng in the October 2021 issue of the Journal of the Royal Statistical Society. Meng argues that it would benefit research for statisticians to treat data as products of research "in and of themselves", rather than as inputs for analysis. Through a rigorous investigation of the entire data lifecycle, we can identify adverse influences on data quality and promote deeper and more explicit clarity from authors about their data, a process called data confession. However, data minding is a lengthy and time-consuming process. No attempt to conduct data minding at scale exists as of this paper, but we aim to change using the novel power of large language models to understand and analyze text. There has been extensive work on the applications of data mining for systemic review and meta-analysis, two close cousins of data minding, but large language models (sometimes referred to as foundation models) like BERT, GPT, and LLAMA, are capable of much more than the established role of automated data processing in these reviews. This paper demonstrates just one small example of their potential to save time and energy for reviewers and data minders. We use a novel sampling method involving semantic similarity between text vectors based on the work of Coecke et al (2010). This allows us to build a meta-dataset made up of research papers across many journals that we can use to fine-tune a foundation model for a specific data minding task, without the need for manual identification. In this case, we were interested in how much time authors devote to each section of the data lifecycle, and we classified text from papers published in *Science* from 1950-2020. We find that while the proportion of papers devoted to methodology has stayed constant across this period, the number of paragraphs concerned with data processing, management, and visualization have grown at the expense of paragraphs discussing data collection and conceptualization and data provenance. We hope that readers will derive two insights from our work. First, from our results, we hope to encourage greater data confession among authors, and a more even distribution of attention paid across the data life cycle in submissions to *Science*. Second, we hope to provide a novel framework for reviewers interested in meta-analysis using foundation models. ChatGPT has already revolutionized the way that industry and the public think about automated tools: this paper shows that academics, specifically reviewers and meta-analysts, shouldn’t overlook the ways in which these models can improve their research.

References

Berry, Michael W., et al., editors. Survey of Text Mining II: Clustering, Classification, and Retrieval. Springer, 2008.

Bommasani, Rishi, et al. On the Opportunities and Risks of Foundation Models. 2021. DOI.org (Datacite), https://doi.org/10.48550/ARXIV.2108.07258.

Meng, Xiao-Li. "Enhancing (Publications on) Data Quality: Deeper Data Minding and Fuller Data Confession." Journal of the Royal Statistical Society Series A: Statistics in Society, vol. 184, no. 4, Oct. 2021, pp. 1161–75. DOI.org (Crossref), https://doi.org/10.1111/rssa.12762.

Pham, Ba', et al. "Text Mining to Support Abstract Screening for Knowledge Syntheses: A Semi-Automated Workflow." Systematic Reviews, vol. 10, no. 1, Dec. 2021, p. 156. DOI.org (Crossref), https://doi.org/10.1186/s13643-021-01700-x.

Vaswani, Ashish, et al. Attention Is All You Need. 2017. DOI.org (Datacite), https://doi.org/10.48550/ARXIV.1706.03762.

Wing, Jeannette M. "The Data Life Cycle." Harvard Data Science Review, June 2019. DOI.org (Crossref), https://doi.org/10.1162/99608f92.e26845b4.