RESEARCHER-IN-THE-LOOP FOR SYSTEMATIC REVIEWING OF TEXT DATABASES

A PREPRINT

Rens van de Schoot
Utrecht University
a.g.j.vandeschoot@uu.nl

Jonathan de Bruin
Utrecht University
j.debruin1@uu.nl

May 28, 2020

ABSTRACT

Machine learning with human interaction is gaining popularity. Humans are intelligent but slow and not very accurate, whereas machines are fast and accurate but not (very) intelligent. Combining the strengths of both in interactive processes can advance the capabilities of machine learning. The idea of combining human intelligence and machine learning can be found in machine learning techniques like active learning (AL) and Human-in-the-Loop (HITL) machine learning. In AL and HITL, the human classifies unlabeled items the model selects to learn from, e.g., items for which the model is uncertain. These techniques are proven to be effective for training models.

In most HITL machine learning applications, the interaction with the human is used to train a model with a minimum number of labeling tasks. However, there is a type of problems where the model is not the primary output, but the systematic exploration of a relevant subset of the given dataset. Usually, these data points should be all be seen, in any case, by a human. In this case, the algorithm(s) are interactively optimized for finding these data points, instead of making the model more accurate. Examples are found in applications were challenges have a unique character and deal with large volumes of data, like systematic reviewing with the goal to create an overview of scientific literature, developing medical guidelines, or police investigations. In these applications, there is also a strong systematic component as all data points need to be judged with the same inclusion criteria. Therefore, we propose the term Researcher-In-The-Loop (RITL) as a special case of HITL with three unique components: (1) The primary output of the process is a selection of the data, not a trained machine learning model, (2) All data points in the relevant selection are seen by a human at the end of the process, (3) The use-case requires a strong systematic way of working.

In this presentation, we study a use-case of RITL machine learning about reading COVID-19 related scientific papers. With the emergence of online publishing, the number of scientific papers on any topic, but especially COVID-19, is skyrocketing. All these textual data present opportunities to scholars and practitioners, while simultaneously confronting them with new challenges like the sparsity of relevant items. For example, this challenge shows up in the development of specific medical guidelines. To develop comprehensive overviews of the relevant studies, scholars often develop systematic reviews. Such reviews entail several explicit and reproducible steps, including identifying all likely relevant publications in a standardized way, extracting data from eligible studies, and synthesizing the results. To avoid bias in systematic reviews, it is of crucial importance to find a way to effectively automatize this screening process. Therefore, we developed software combining natural language processing and active learning: *ASReview: Active Learning for Systematic Reviews*. The source code is available open source via GitHub. We show that by using ASReview the labor-intensive task of labeling text leads to far more efficient reviewing than a full manual review.

***K*eywords** Researcher-in-the-Loop · Human-in-the-Loop · Active Learning · Machine Learning · Interaction · Natural Language Processing · Deep Learning · Artificial Intelligence