Hybrid Data Annotation Models for AI-Prototyping

Making Rule-based Systems and Supervised Learning Algorithms work together.

Prof. Falk Uebernickel*, Prof. Bernhard Renard, Prof. Gerard de Melo

A joint project between the chairs of Design Thinking and Innovation Research, Data Analytics and Computational Statistics, and Artificial intelligence and Intelligent Systems
This master project is offered exclusively to DATA ENGINEERING master students!

Problem description and project goal

For Supervised Learning (SL) use cases that do not already have large amounts of (historically) annotated data available, generating and owning qualitative training data is a significant entry barrier. Without such an existing data set, the annotation is labor-intensive and often requires week-long repetitive work. This master project aims to design and implement a process that shortens the time needed to annotate data for SL projects using a hybrid approach where the amount of work for manually labeling data sets is reduced.

How do we imagine the hybrid approach going to work? In a “naive” annotation procedure, there is no differentiation between explicit and tacit knowledge. We want to make use of this differentiation by providing the possibility to list explicit domain knowledge via a set of heuristics, which is used to annotate ideally large parts of the data automatically. Furthermore, the manual data annotation process can be enhanced by training SL models on the data by prioritizing the least confident data.

Implications would include a much faster annotation process for such use cases that can be described using these sets of heuristics. Thus, this would also enable AI prototyping for idea-driven use cases instead of just data-driven use cases. Furthermore, this technology will reduce data annotation costs. Lastly, the explicitly defined domain knowledge can be used as documentation to explain model decisions better.

Work packages

1. Development of a baseline system (annotation service and heuristics engine)
2. Enhancement of the heuristics engine (heuristics exploration and detection engine, heuristics correctness, weighted heuristics)
3. Development of an AutoML system to enhance manual data annotation
4. Evaluation of prediction quality using standard metrics and visualization techniques

Requirements

- Willingness to annotate larger amounts of data as part of the project
- Experience from working in a startup as an entrepreneurial and innovative mindset is required
- Profound knowledge of ...
    - Flask (Backend technology),
    - MongoDB and Postgres (Database systems),
    - Angular and D3.js (Frontend technologies),
    - Snorkel, Scikit Learn, and PyTorch (Machine Learning technologies)

Contact

Prof. Dr. Falk Uebernickel
Email: falk.uebernickel@hpi.de

* project lead