15th European DDI Users Conference, Ljubljana

EDDI2023

Contribution ID: **50**　　　Type: **Regular Presentation**

Survey Variables Classification with Hierarchical Machine Learning

Wednesday, 29 November 2023 11:30 (30 minutes)

Recent developments in Machine Learning (ML) show robust performance in the area of Natural Language Processing (NLP) tasks, such as sentiment analysis and document classification. Our ML task is one of short text classification, specifically we are endeavouring to annotate variables using the variable name, label, question text and representation. Our task is one of multi-class classification, where accuracy is known to be sensitive to the number of labels in the dictionary. For this particular ML task we are bounding the annotation to purely learn 'key variables'- socio-demographic indicators which feed our Disclosure Risk Analysis (DRA).

In this presentation, we present a Hierarchical Machine Learning (HML) approach to recognising variable concepts from studies available at the UKDS. Specifically, we decompose the task into first learning the broad group to which the variable belongs, e.g. Education and then the concept within that group, e.g. Highest Educational Qualification.

At a high level, we use a mix of shallow and deep learning models to minimize the number of target class labels and boost the overall performance of each individual algorithm. We present details of the cost function simplification process and each hierarchy metrics and benchmarks.

Primary author: EVDOKIMOV, Ivan (University of Essex)

Presenter: EVDOKIMOV, Ivan (University of Essex)

Session Classification: DDI - New Directions

Track Classification: Proposed Topics of the Conference: Machine Learning, AI and Automation