Keynote 3

Challenges in explaining machine learning models for text

Marko Robnik Šikonja
University of Ljubljana, Faculty of Computer and Information Science
Ljubljana, Slovenia
marko.robnik@fri.uni-lj.si

Abstract

The area of Explainable Artificial Intelligence has developed many approaches for the explanation of machine learning models. The most successful methods are based on counterfactuals, prototypes, and perturbation of inputs. Unfortunately, none of these approaches works well to explain large language models, which currently dominate natural language processing. The presentation will focus on challenges in using the most successful explanation methods for text classification, such as the interpretation of feature attributions, explanation of longer textual units, on-manifold vs. off-manifold explanations, unstable and uncertain explanations, and inadequate and unsystematic evaluation of explanation techniques. We will present possible solutions and outline a framework for more general explanation approaches.