Bachelor/Master Thesis

Karlsruher Institut für Technologie

Institut für Programmstrukturen und Datenorganisation (IPD)
Lehrstuhl für Systeme der Informationsverwaltung, Prof. Böhm

Bachelor Thesis

Tuning von Explainable Artificial Intelligence (XAI) Werkzeugen für die Textanalyse

(*engl.: Tuning of Explainable Artificial Intelligence (XAI) tools in the field of text analysis*)

Explainable Artificial Intelligence (XAI) comprises methods, so-called *explainers*, to explain Machine Learning (ML) models. Existing explainers support many real-world ML applications, such as image and text classification. One specific approach is Post-Hoc explanation that explains black-box ML models. They do so by discovering which features are important for the predictions. For example, a Post-Hoc explanation may discover that certain words in a text have a strong impact on the text classification.

Nowadays, different Post-Hoc explainers exist, and it is important to know how *well* such explainers work. We try to answer this question as part of our current research. This also includes providing end users with guidelines on the correct use of explainers, in order to receive *good* explanations. We expect that such guidelines eventually improve the usefulness of explanations. An example of such a guideline in the context of text classification is to emphasize that a certain text hierarchy in a text corpus, e.g. paragraphs, provides a *good* explanation. So the guideline might instruct the end user to rather ask the explainer for paragraphs to explain the respective text classification, instead of words or sentences.

Our use case, explanation of sentiment classification, can be divided into three steps. First, one classifies a text to predict a sentiment. Second, one sets the parameters of the explainer, such as the text hierarchy, e.g., paragraphs or words. Third, the explainer provides an explanation. Here, a possible explanation would be to highlight the part of the text that is most relevant for the classification.

In this thesis, you will focus on the second step by deploying specific Post-Hoc explainers and analyzing their behavior. Based on the outcome, you will design evaluation metrics to measure how well the explainer works for a particular text hierarchy. Your ultimate goal then is to draw conclusions regarding the general usage of the respective Post-Hoc explainer.

This results in the following specific tasks for this thesis:

- Literature research on explainers for sentiment classifiers (e.g., SHAP [1])
- Extension of an existing framework for XAI evaluation with selected explainers
- Evaluation of the explanations generated, and deriving guidelines on how to use explainers

XAI is a hot topic in the machine learning community and society. Your results directly contribute to research in this field and advance the state-of-the-art of XAI. In the context of this thesis you have the possibility to work with latest technology, methods and tools in the area of machine learning and data science, e.g. the LIME and SHAP frameworks, scikit-learn and word embeddings. For experimental evaluation, you can rely on our institute infrastructure with distributed computing clusters and dedicated GPU servers.
In the context of this thesis it is advantageous to know the basics of data analysis and to have a fundamental knowledge of programming languages like Python.

[1] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* (Vol. 2017-December, pp. 4766–4775). Neural information processing systems foundation.

Kontakt

Clemens Müssener clemens@muessener.com

Am Fasanengarten 5 76131 Karlsruhe Building: 50.34