The Institute for Ethics in Artificial Intelligence (directed by Prof. Dr. Christoph Lütge) offers the following thesis topic (Master):

Explainable AI (XAI) in Natural Language Processing (NLP)

Background

Natural language processing (NLP) is a subfield of artificial intelligence and focuses on programming computers to understand and generate human language. At Munich Re, we use NLP to solve use cases such as extracting relevant data for underwriting and claims handling. A key method in NLP are language models (LM) which generate probability distributions over sequences of words to predict, e.g., the next word in a sentence. New generation LMs, such as BERT, GPT-3 and Switch Transformers, are deep learning models with millions, billions or even trillions of parameters trained on hundreds of gigabytes of text. These models are able to capture a superior language understanding including subtle linguistic details which render them capable of adapting to new NLP tasks with little or even no additional data.

However, the models complexity essentially renders them black box systems which brings along reduced trust from the user side. Additionally, such models may fail in counterintuitive ways and are opaque in their decision-making process. This raises a need for explainability.

Goal

The goal of this master thesis is to provide a background on interpretation techniques, i.e., methods for explaining the predictions of NLP models. The thesis should provide a literature review on XAI methods for NLP including their respective pro's and con's in view of Munich Re's activities. This shall be the basis for a high-level high level guideline on when to use which method in terms of type of the problem and nature of the use case, for example. Selected methods shall be implemented in Munich Re's proprietary NLP package in order to showcase the application on an existing use case. Potential open-source packages to be worked with are Captum, Alibi, the Adversarial Robustness Toolbox and TextAttack, for

example.

Profile

- Quantitative background in computer science, mathematics, natural sciences or similar with an interest in explainable AI.
- Very good Python coding skills.
- Ideally, familiarity with PyTorch, Tensorflow and interpretability packages (e.g., Captum).

Contact

If you are interested, please contact (ieai@mcts.tum.de)

We are looking forward to your application!

References

- Danilevsky, Marina, et al. "A survey of the state of explainable AI for natural language processing." arXiv preprint arXiv:2010.00711 (2020).Very good Python coding skills.
- https://captum.ai/
- https://github.com/SeldonIO/alibi
- https://github.com/QData/TextAttack