Explainability in AI - The Case of Burnout Detection

Degree programme : BSc in Computer Science | Specialisation : Data Engineering
Thesis advisor : Prof. Dr. Mascha Kurpicz-Briki

Predicting illnesses or syndromes using Artificial Intelligence (AI) often suffers from a lack of trust in the clinical environment. This thesis investigates the state-of-the-art techniques available in the evolving field of explainable AI, to not only improve trust in these decisions, but also show how data engineers and data scientists can use it to avoid common pitfalls in AI technology.

Introduction

Burnout, a state of emotional, physical, and mental exhaustion caused by excessive and prolonged stress, is a growing concern in our working environments. Previous work in the field has shown, that Artificial Intelligence (AI) can be used as an effective means to detect indicators for burnout in free text samples, collected on the social media platform Reddit. However, these predictions often suffer from a lack of trust in the clinical environment. State-of-the-art techniques available in the evolving field of explainable AI provide an interesting approach to close this gap.

Research Questions

To determine if explainable AI is suitable for this task, we answer the following research questions:

- What explainability techniques can be used for Natural Language Processing (NLP)?
- How can they help to improve trust in AI?
- How does this also help detecting and avoiding common pitfalls?

State-of-the-art

The trade-off between the accuracy and interpretability for AI models usually leads to highly accurate models being preferred, but they are inherently complex to explain, even for seasoned professionals. A literature review provides a summary over the properties and approaches to explain such black-box models. Most notably in the field of NLP is the usage of surrogate models. With this method, a simpler and more interpretable model is derived to directly build explanations that are easy to understand by humans. Using surrogate models in combination with other techniques, such as feature importance, many highly advanced libraries shape todays standard and allow for visualized explanation to be generated, both for specific examples, as well as for the entirety of an AI model.

Result

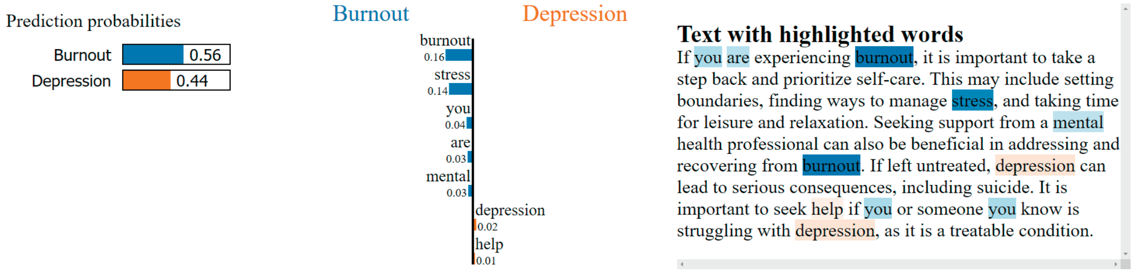
Three popular libraries, trying to tackle the issue of explainable AI, stand out:

- Local Interpretable Model-agnostic Explanations (LIME)
- SHapley Additive exPlanations (SHAP)
- Explain Like I'm 5 (ELI5)

The implementation and exploration of these libraries on an existing dataset for burnout detection from previous research and with respect to the research questions, show that reasonable and reliably surrogate models can be derived using easy-to-use and well documented high-level APIs. The provided explanations and visualizations, especially in the example-driven field of NLP, are simple to understand for people, even without clinical or technical background. Information gained from explainable AI can also be used by data engineers and data scientists to further improve their AI models. Despite all the advantages that explainable AI brings, a proper evaluation of the libraries, techniques, as well as their outcome, is however an absolute necessity to avoid further pitfalls.

Sascha Beat Schwärzler

sascha.schwaerzler@
blackyeti.ch



Text with highlighted words

If you are experiencing burnout, it is important to take a step back and prioritize self-care. This may include setting boundaries, finding ways to manage stress, and taking time for leisure and relaxation. Seeking support from a mental health professional can also be beneficial in addressing and recovering from burnout. If left untreated, depression can lead to serious consequences, including suicide. It is important to seek help if you or someone you know is struggling with depression, as it is a treatable condition.

Hard decision – Popular explanation model LIME providing reasoning on why the given text was classified as "Burnout"