Detecting Indication for Anxiety in Free-Text

Degree programme : BSc in Computer Science | Specialisation : Data Engineering
Thesis advisor : Prof. Dr. Mascha Kurpicz-Briki
Expert : Han van der Kleij (SBB CFF FFS)

Anxiety disorder is one of the most common mental illnesses, affecting 15-20% of the Swiss population according to the University Hospital Zurich. However, the diagnosis is challenging. This thesis explores the potential of Natural Language Processing (NLP) techniques in detecting anxiety disorders, aiming to assist clinical professionals in their work.

Goal

This thesis aims to investigate the potential of utilizing machine learning techniques, particularly NLP, for anxiety detection in free-text data. Furthermore, the objective is to highlight the distinctions between anxiety-related and control (not anxiety-related) data.

Data

In a pre-project, posts from various anxiety-related and not anxiety-related (as a control group) forums of the social news platform Reddit were crawled, anonymized and partially labelled by human annotators. Using this data as basis, multiple datasets were created consisting of either manually labelled anxiety and control data, or automatically labelled anxiety and control data.

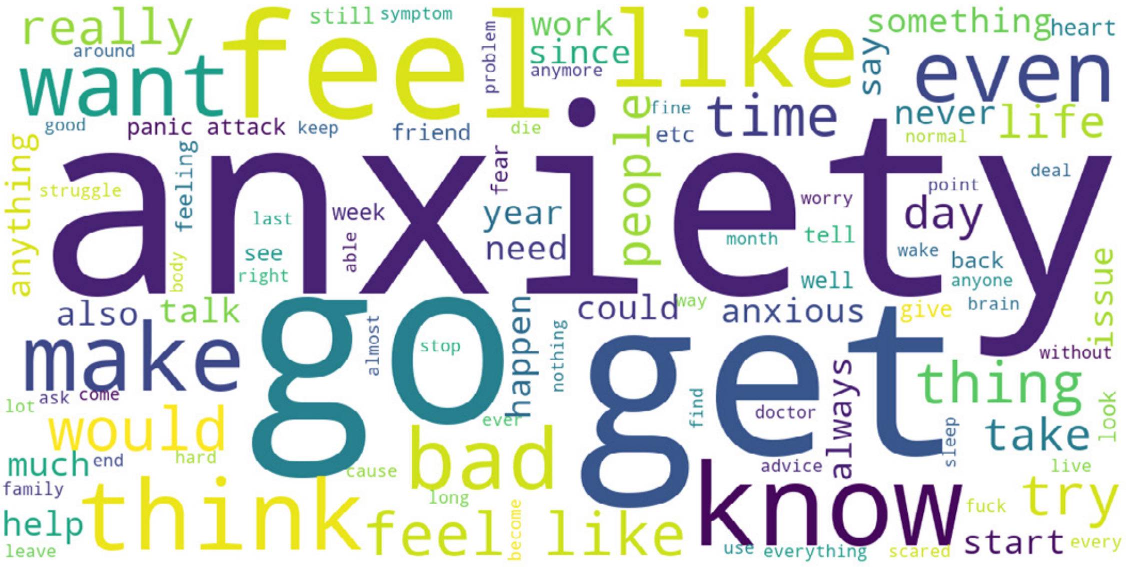
Methods

The different data sources and defined datasets were analyzed to identify initial differences. The posts were then preprocessed using techniques like lemmatization, lowercasing and stopword removal. Using a Token Frequency - Inverse Document Frequency (TF-IDF) approach as a feature generation technique, the different datasets were used as input data for a variety of classifiers. Experimenting with 120, 300 and 500 features as well as feature removal, the results were measured using the metrics accuracy and recall. As an additional experiment, the possibilities of using a custom ensemble model were explored.

Results

Sentiment analysis based on the used words revealed that the language in the anxiety-related samples is more negative than in the control samples. Among the tested single classifier models, SVM with the most used 500 tokens as features demonstrated the highest performance across all datasets, achieving an accuracy of 95.7%. Removing the most important feature token, „anxiety," the models performance only decreased around 3-4%. Although the created ensemble models performed better in finding all relevant anxiety posts (high recall), their overall accuracy was lower compared to the single classifier models.

Tobias Kocher



Word cloud showing the most commonly used words in the human-labelled anxiety data group.