Burnout Detection in Turkish Text

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

Burnout Text in Turkish Text is a natural language processing (NLP) text analysis project using supervised machine learning (ML). The aim of this project is to detect burnout syndrome in a given text. Psychological problems are increasing in society worldwide. Therefore, this project aims to develop a tool to help specialists in the field of psychology and psychiatry.

Motivation

Psychological problems are on the rise worldwide. The Covid-19 pandemic has caused many unusual changes in our lives. After the Covid-19 pandemic, a further increase in psychological problems has been observed worldwide. Because of the increase in psychological problems, more specialists and techniques are needed in this field. This project was considered as an application to assist psychiatrists and psychologists in their work.
Experts ask open-ended questions to a person who may be suffering from burnout syndrome. The written answers to these questions can be analyzed with ML and burnout syndrome can be detected.

In the past years, machine learning has made great progress. There are three main reasons for this progress. I list below three main reasons that create a favorable environment for machine learning projects.
- Humans generating more data as a result of digitization
- Increased computing power of computers
- Success of algorithms used in machine learning.

We can also use the progress in machine learning to detect burnout syndrome. The texts output from open-ended burnout-related questions can be analyzed using machine learning. This is the main goal of the project. Our goal is to classify a given text regarding burnout syndrome.

Content of the Thesis

This project is an NLP (Natural Language Processing) project. The data used for the project was obtained from an internet forum called Eksi Sözlük. Two categories of data were used for the project. The first category is burnout syndrome entries. The second category of data is the so-called neutral data. The neutral data category is data selected from various topics such as „sun“ „evolution“ „Cern“ „Game of Thrones“.
Supervised machine learning is used in this project. For this, the data needs to be annotated. The data was annotated by two independent annotators.

Four categories of data sets were prepared:

- Non-Balanced & Non-Lemmatized Dataset
- Balanced & Non-Lemmatized Dataset
- Non-Balanced & Lemmatized Dataset
- Balanced & Lemmatized Dataset

Logistic Regression, Random Forest, Support Vector Machine and AdaBoost algorithms were used in the project. In addition, CountVectorizer and TD-IDF vectorization techniques were used to vectorize the data sets.

Result

As a result of the experiments and investigations, the best results were obtained with balanced data sets. The application of the lemmatization process to the dataset increased the score obtained.
The TF-IDF vectorization technique also led to a significant increase in the score of some of the models. Therefore, the best result was obtained by applying the TF-IDF vectorization technique on the lemmatized balanced dataset. Very good results were obtained by applying TF-IDF vectorization technique on lemmatized balanced dataset in Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), AdaBoost (AB) algorithms.

The best result with **0.92** recall score was obtained with algorithms Logistic Regression and Support Vector Machine. 10-Fold CV method was applied in the models. This avoids the biasing effect that having little data can have on the model. It is aimed to test the model with real clinical data in the future.

Barin Kaya