CSSM502 Advanced Data Analysis in Python
Mehmet Fuat Kına
12/14/2025

Midterm Project Proposal

Cross-Linguistic Patterns of Work-Related Stress on Reddit

People experience and describe work-related stress in different ways, even when they face similar working conditions. While psychological theories such as Lazarus' Cognitive Appraisal Theory offer conceptual explanations for these differences, most existing studies rely on surveys or predefined categories rather than naturally occurring language. My project takes a computational approach to examine how work-related stress is structured in large-scale online text, without imposing theoretical categories in advance. The main objective is to identify latent semantic and thematic structures that organize stress-related narratives and to examine how these structures vary across different social contexts.

The data for this study will come from public Reddit entries about work and burnout (from subreddits such as burnout, jobs, and work). Data collection is planned through the official Reddit API using Python. An API request has already been submitted, and approval is still pending. Then, the text will be prepared using a standard preprocessing steps. In addition to English, the project also plans to include data from Turkish. Including more than one language makes it possible to examine whether patterns in stress-related discourse stay similar across languages or whether they change depending on linguistic context.

From a methodological point of view, the project relies mainly on representation learning and unsupervised analysis. Each post will be converted into a numerical representation using pre-trained multilingual sentence embedding models. These models allow texts from different languages to be mapped into the same semantic space, which means that comparison is based on meaning rather than on specific words. After this step, unsupervised methods such as dimensionality reduction and clustering will be applied to explore whether there are recurring patterns in how work-related stress is discussed. Because unsupervised methods can sometimes produce unstable results, the analysis will also check whether the identified patterns remain similar across different random initializations, subsets of the data, and languages. Psychological theory is not used as a fixed structure that the data must fit into. Instead, Lazarus' appraisal categories, such as threat, coping, and challenge, will be used after the main patterns are identified. The goal here is to see whether the patterns found in the data resemble these appraisal dimensions or whether they point to something different. If some patterns do not align well with existing theory, this will not be treated as a problem but as an interesting result. In this sense, the project also looks at the limits of how well psychological theories of stress apply to naturally occurring language, especially across different languages.

The analysis will begin with simple descriptive statistics and exploratory analysis to get a general sense of the data. This will be followed by unsupervised modeling and cross-linguistic comparison of the discovered patterns. The main focus of the project is not on prediction accuracy but on interpretation and theoretical relevance. By combining multilingual text representations with data-driven analysis, the project aims to better understand how work-related stress is structured in online narratives and how existing psychological theories relate to these structures.