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

Midterm Project Proposal

Comparing Theory-Driven and Data-Driven Text Representations of Work Stress using Reddit Data

Lazarus's Cognitive Appraisal Theory argues that stress depends on how individuals evaluate a situation and their ability to cope with it. According to Lazarus's perspective, a stressful experience can be viewed in different ways. The aim of this project is to investigate how work-related stress is expressed and appraised in online narratives by analyzing written posts in which individuals describe their work experiences in their own words.

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.

A central methodological challenge of the project is to translate Lazarus' appraisal categories into features that can be analyzed computationally. Threat appraisal is expected to appear in language expressing fear, anxiety, uncertainty, or loss. In the data, different types of appraisal are expected to appear through the way people describe their experiences. Posts reflecting coping appraisal often include attempts to ask for advice, deal with a problem, or manage emotions. Challenge appraisal, is more likely to appear in posts where users talk about learning from difficulties or trying to see a stressful situation in a more positive way. These categories are offer a useful starting point for analysis.

Methodologically, each post will be represented using two alternative text representations. First, a theory-driven representation will be constructed using dictionary-based features aligned with threat, coping, and challenge appraisals, operationalized as frequency-based scores. Second, a data-driven representation will be created using TF–IDF features derived directly from the text. Using the same dataset and analytical framework, these two representations will be compared to assess whether psychologically informed features provide additional explanatory value beyond standard text representations.

The analysis will begin with descriptive statistics and exploratory text analysis to examine general patterns in the data. Unsupervised methods such as clustering will be used to explore whether posts naturally group according to appraisal-related features. In addition, simple supervised models will be applied using the same modeling setup across both feature representations to support direct comparison. In the supervised analysis, appraisal outcomes will be operationalized as continuous scores derived from relative word frequencies, and where necessary simplified using transparent thresholding strategies. Clustering results will be interpreted by examining their alignment with appraisal score distributions, allowing for assessment of whether unsupervised groupings reflect theoretically expected appraisal patterns.Rather than focusing on prediction accuracy, the analysis will emphasize interpretability, feature contribution, and consistency with Lazarus' theoretical framework.