Exploring Work-Related Stress Through Unsupervised Text Analysis on Reddit

CSSM502 Final Project

Şevval Çakıcı | Koç University

01.21.2026

Contents

1. Research Purpose & Hypothesis

People experience and describe work related stress in different ways, even when they face similar working conditions. Work related stress is a major issue today. People experience and talk about it differently depending on their social and cultural backgrounds. Psychological theories like Lazarus Cognitive Appraisal Theory offer structured ways to explain stress. Stress has a solid scientific background but most research uses surveys or set categories. These methods can miss how people naturally express stress in their daily language.

Recent advances in computational social science make it possible to analyze large scale text data. These naturally occurring data are important to study social phenomena beyond predefined theoretical constructs. Rather than asking individuals to report stress through fixed questionnaires, online platforms allow researchers to observe how people spontaneously describe work experiences, frustration, and economic pressure in their own words. This project takes a exploratory approach to examine how people discuss work related stress online. The study looks for hidden patterns in large amounts of text without setting categories ahead of time, by using representation learning and unsupervised methods. Reddit is a good place for this research because it has many user discussions where people share their thoughts on work, burnout, money worries, and trust in institutions.

The project first aimed to compare English and Turkish discussions using multilingual sentence embeddings. However, problems with the Turkish data led to a focus on English-language Reddit posts. Although this means a direct language comparison was not possible, the methods used can still be applied to other languages and allow for a detailed look at English data.

With this motivation, the study follows exploratory expectations instead of strict, testable hypotheses. First, the study expects that discussions about work related stress are not all the same. Instead, they are likely to form several distinct groups that reflect different aspects of work, such as economic insecurity, workplace relationships, trust in institutions, and emotional exhaustion. Second, the study expects that these groups will only partly match established psychological theories of stress, like Lazarus Cognitive Appraisal Theory. This suggests that everyday language captures parts of stress that go beyond set theoretical categories. Third, the study expects that unsupervised text analysis can uncover hidden patterns in how people describe and understand work-related stress. It also aims to show the limits of using only data driven methods to capture personal experiences.

The main goal of this study is to find common patterns in how people talk about work related stress and see if these match or differ from existing psychological and sociological ideas. The focus is on exploring and interpreting the data, not on prediction accuracy, to show both the strengths and limits of unsupervised text analysis in social science.

2. Analytical Model

2.1.Data Preparation

The analysis uses English language Reddit comments from publicly available datasets on Kaggle. These datasets collect user posts from several subreddits where people often discuss work, employment, burnout, and economic issues. Using these secondary data sources was necessary because direct data collection through Reddit's API and other services was limited, so public datasets were the best option for large scale analysis.

An initial pool of about 64,000 comments was created by combining data from several relevant subreddits. This pool was cleaned to improve data quality and consistency by removing duplicates, very short texts, and non-linguistic content. After cleaning, 59,895 comments still remained. To keep the analysis manageable and maintain topic variety, a random sample of 10,000 comments was taken from the cleaned dataset. Random sampling was used instead of keyword filtering to avoid reinforcing fixed ideas about stress and to allow for a wider range of expressions of work-related experiences.

As part of the original research design, a Turkish-language dataset was also created using keyword-based filtering on a publicly available Reddit corpus (n = 1,753). Manual review showed a lot of unrelated content such as profanity, pop culture references, and isolated economic terms, without a broader work-related context. Due to its small size and lack of thematic focus, the Turkish dataset was excluded from the final analysis. The aim is to maintain validity. This situation has prevented direct comparison between languages. But I believe it will allow for a more reliable examination of English discourse. It will also leave room for future research in other languages.

Preprocessing was kept minimal to avoid removing important language differences. No stemming or heavy stopword removal was used, since sentence-level embedding models are meant to capture meaning from full natural language input.

2.2.Text Representation

Each comment was encoded as a fixed length numerical vector using a pre trained Sentence-BERT model. The texts were represented as 384 dimensional embeddings, such that comments with similar semantic content are located closer to one another in the embedding space, even when they do not share the same vocabulary.

This approach takes the analysis further than simple keyword matching and focuses on understanding meaning within its context. This is particularly useful when studying work-related stress, because people do not always describe stress in a direct or clear way. Instead, they often express stress through personal stories, irony, or criticism of institutions, rather than by using explicit stress-related terms. By comparing the meanings of sentences, the embedding model helps uncover less visible patterns in how people talk about their work experiences and stress in everyday language.

2.3.Dimensionality Reduction

To explore latent semantic structure in the embedded text data, unsupervised dimensionality reduction and clustering techniques were applied. Since the original embedding space is high-dimensional, dimensionality reduction was used primarily for visualization rather than for altering the underlying representations.

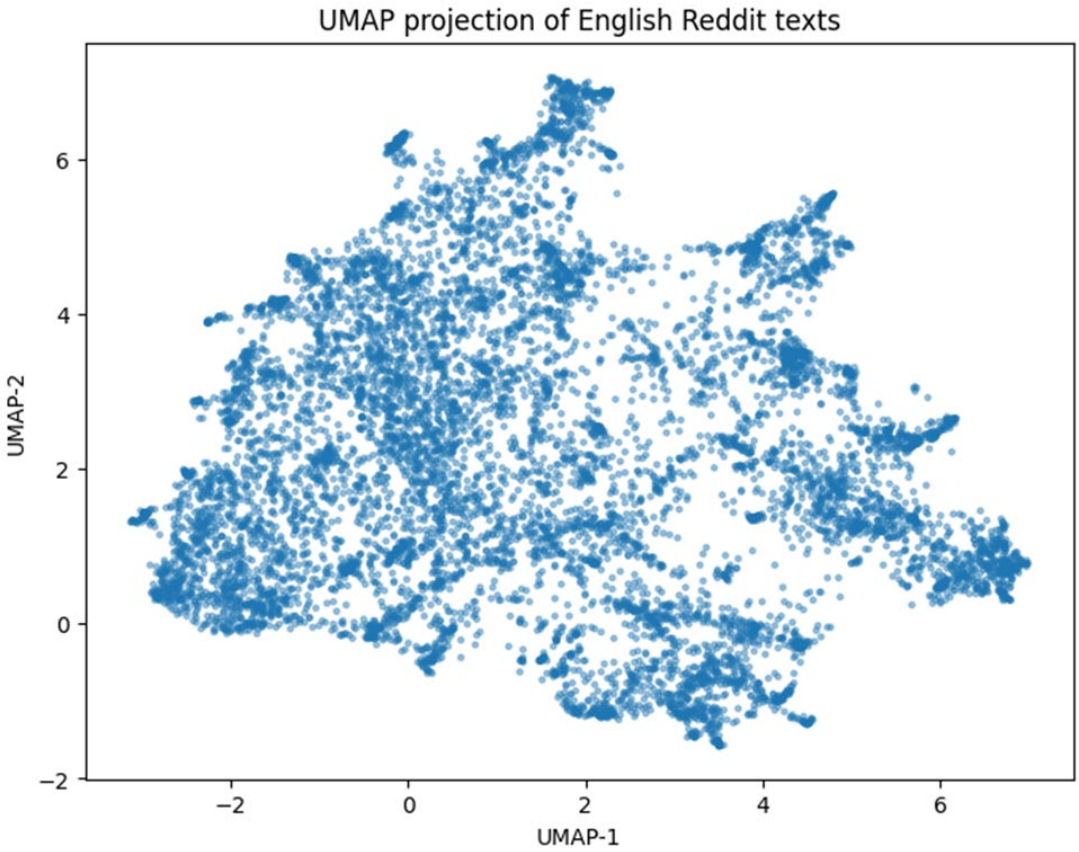

Figure 1. *UMAP projection of 10,000 English Reddit comment embeddings. Each point represents a comment encoded with Sentence-BERT. The visualization shows multiple locally dense regions not single cluster.*

Uniform Manifold Approximation and Projection (UMAP) was employed to project thet 384 dimensional embeddings into a two-dimensional space. UMAP was chosen due to its ability to preserve local neighborhood structure. UMAP makes it well-suited for identifying dense semantic regions in text data. The resulting visualization revealed multiple locally dense areas. It does not reveal a single homogeneous cluster. Visualization suggests the presence of thematic patterns.

2.4.Clustering

Clustering was carried out with the KMeans algorithm using the original embedding vectors rather than the reduced UMAP space. This choice ensures that the clustering is based on the full semantic information contained in the embeddings. After several exploratory tests, the number of clusters was set to eight in order to keep the results interpretable while still capturing sufficient detail.

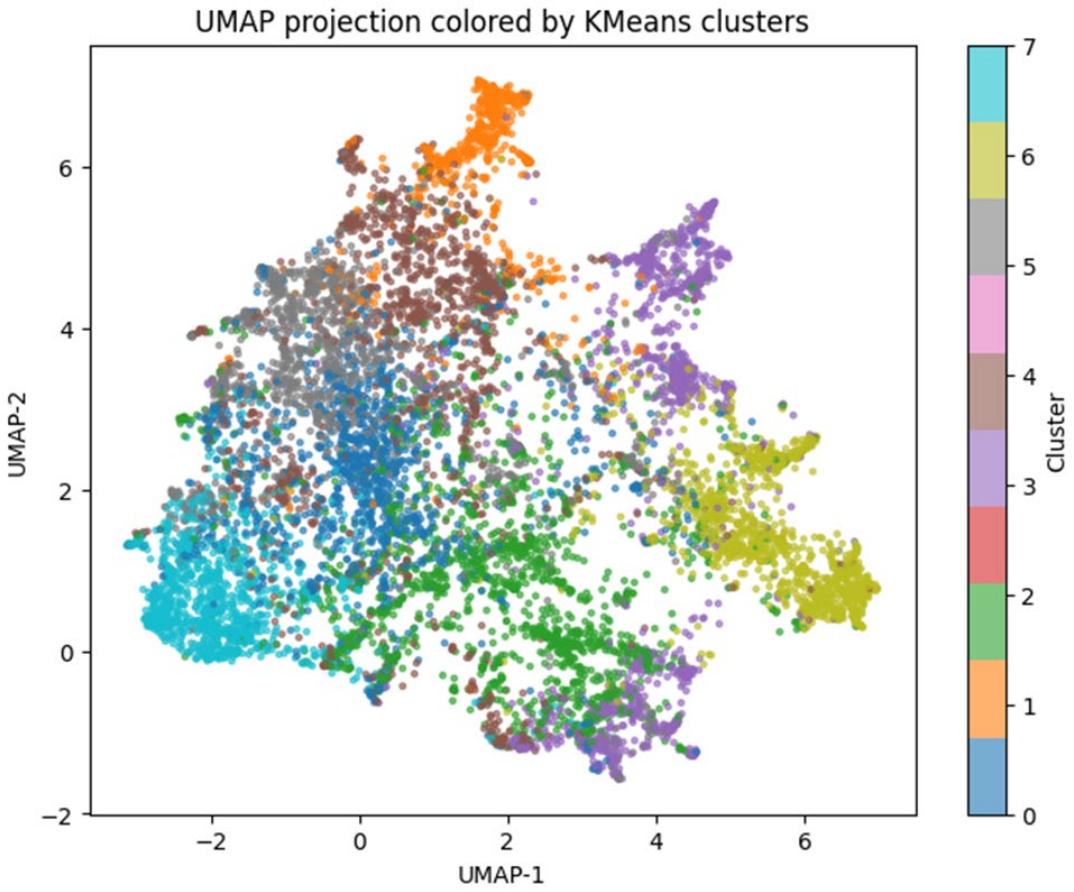

Figure 2. *UMAP projection colored by KMeans cluster assignments (k = 8). Some clusters form compact and well-separated regions. Others exhibit partial overlap. It reflects the fuzzy boundaries of work related discourse themes.*

The resulting clusters varied in size but were all large enough for qualitative inspection. Sample texts from each cluster showed clear thematic groupings, such as workplace frustration, political and institutional critique, gaming-related stress, relationship advice, and cultural commentary. This diversity supports using unsupervised methods to discover naturally emerging discourse categories without predefined labels.

3. Findings

This section presents the empirical findings. These are obtained from the unsupervised analysis of English-language Reddit comments. The results are organized around the structure and content of the discovered clusters. Focus is on identifying work related discourse pterns and semantic diversity.

3.1. Cluster Overview

The unsupervised clustering procedure produced 8 distinct clusters based on semantic similarity in the sentence embedding space. Although the sizes of these clusters varied, they were all large enough to allow for qualitative analysis. Visualization of the clusters using UMAP

revealed multiple locally dense regions with partial overlap rather than sharply separated groups, indicating that discourse related to work and stress does not form rigid boundaries but instead consists of interconnected thematic areas.

Looking across the clusters, a consistent sense of semantic coherence was apparent when representative examples were reviewed manually. Some clusters appeared as relatively tight groupings in the embedding space, whereas others overlapped with one another, reflecting the fact that online discourse does not follow clear boundaries and often varies by situation. This overall arrangement shows that unsupervised approaches are suitable for finding meaningful patterns, without pushing the data into fixed thematic categories.

3.2. Identification of Work-Related Clusters

To identify clusters linked to work and stress, representative texts from each cluster were reviewed qualitatively. This reading was supported by simple keyword-based checks. Around 86.5% of all texts contained at least one term related to work, the economy, or stress. At the same time, the level of topical noise was estimated to remain below 20%. These checks were used only as a supporting tool. They did not serve as the main criterion for identifying or defining the clusters.

Based on thematic coherence and a close reading of the content, three clusters (Clusters 0, 4, and 5) were identified as the most closely related to work. These clusters showed a high presence of work and economy related terms. During manual review, they also displayed a stable and shared focus around work related themes.

Cluster 0 is more about dealing with attitudes toward the workplace and everyday job experiences. Many texts mention dissatisfaction with working conditions, management, or job security.

Cluster 4 is more about institutions and wider social frustration. Posts in this cluster often refer to political systems, labor markets, and what users see as structural unfairness.

Cluster 5 mainly concerns economic inequality and material insecurity. It brings together discussions of financial pressure, rising living costs, and frustration with unequal economic outcomes.

Clusters mainly linked to entertainment, gaming stress, or relationship advice were not included in the next stage of the analysis, because their content was not directly connected to work or economic experiences.

3.3. Semantic Diversity of Work Related Stress

Analysis of the retained work-related clusters highlights the semantic diversity of how work-related stress is articulated in online discourse. Stress is not shown only through clear mentions of exhaustion or burnout. Instead, it appears in different ways, such as stories about unfair treatment, lack of trust in institutions, economic complaints, humor, sarcasm, and personal reflections.

The UMAP visualization of the final work-related set (n = 3,330) shows this variety more clearly. Instead of one single cluster, the projection displays several overlapping areas of meaning. This suggests that work-related stress is composed of interconnected issues rather than a single theme.

These findings suggest that stress-related experiences are often embedded within broader social and economic narratives. As a result, methods that depend too much on clear stress related words or fixed survey categories may fail to capture how common and complex work related stress is in day-to-day language.

4. Theoretical Interpretation

The findings of this study can be linked to existing psychological and sociological theories of stress. At the same time, they point to the limits of applying fixed theoretical frameworks to language that emerges naturally. Instead of treating theory as a strict system of classification, this study follows a post-hoc interpretive approach. Theory is used as a lens to place the patterns in context, after they emerge from the unsupervised analysis.

Lazarus' Cognitive Appraisal Theory provides a useful reference point for interpreting some of the discovered clusters. Several themes identified in the work-related clusters resemble core appraisal dimensions such as perceived threat, loss, and challenges to coping resources. For example, narratives centered on economic insecurity and material precarity echo appraisals of threat and loss, while discussions of workplace dissatisfaction and unfair treatment often reflect concerns related to control, agency, and coping capacity. In this sense, parts of the discourse align with established stress appraisal concepts.

At the same time, the match between data based clusters and theoretical categories is only partial. Many expressions of work-related stress do not fit clearly into specific appraisal dimensions. Instead stress is often expressed through wider social stories. These patterns suggest that everyday expressions of stress often blend personal appraisal with collective and structural concerns, extending beyond the individual level focus emphasized in many psychological models. This partial match has important implications. It does not mean that the theory is wrong. Instead, it shows that stress depends on context and how people experience and share it in digital spaces. Online discussions allow people to describe stress not only as a personal feeling, but also as a shared social experience shaped by economic and institutional conditions. As a result, stress related language may reflect grievances, identities, and value judgments, that are not easily captured by survey-based measures or predefined response categories.

From a methodological point of view, these findings show both the advantages and the limits of using unsupervised text analysis in social science research. Unsupervised methods are useful because they can reveal hidden semantic patterns without relying on strong theoretical assumptions from the beginning. For this reason, they are especially helpful when studying complex and multi-layered issues, such as work-related stress. At the same time, interpreting the results still requires theoretical judgment. The boundaries between clusters are not sharp, but instead appear blurred. Instead of producing strict categories, meaning develops gradually. It

becomes visible through careful and contextual reading, rather than through fixed or sharply defined classifications.

Overall, the theoretical interpretation points to a complementary relationship between data-driven and theory-driven approaches. Data-driven methods do not replace psychological theory. Instead, they help extend its reach.

References

Lazarus, R. S. (1974). Psychological stress and coping in adaptation and illness. *The International journal of psychiatry in medicine*, *5*(4), 321-333.