Tilburg University

USING UNEMPLOYMENT AND EDUCATION TO PREDICT MENTAL HEALTH OUTCOMES

EVIDENCE FROM UNDERSTANDING SOCIETY AND CFPS

WANGBO TAO

thesis proposal
data science & society

student number

u305684

committee

Dr. Federico Zamberlan

location

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

date

September 3rd, 2025

word count

825

USING UNEMPLOYMENT AND EDUCATION TO PREDICT MENTAL HEALTH OUTCOMES

EVIDENCE FROM UNDERSTANDING SOCIETY AND CFPS

wangbo tao

1 project definition, motivation & relevance

Since the Covid-19 pandemic and the recent wave of artificial intelligence, unemployment has become a major concern in many societies, especially for young graduates and middle-aged workers. In the European Union and the Eurozone, youth unemployment rates remain high at more than 13–14%. In the United Kingdom, youth unemployment is around 14%, while in China, the youth unemployment rate reached a record 21.3% in 2023.

While many previous studies have shown that unemployment and education are related to mental health, fewer have looked at the predictive value of more detailed unemployment features, beyond a simple unemployment indicator. Even fewer studies have compared these patterns across different labor markets and cultural contexts.

This Thesis will investigate to what extent unemployment status and educational attainment—together with detailed features of unemployment histories (such as duration, frequency, and age at first unemployment)—predict individual mental health outcomes in a cross-sectional setting. The focus will be on two large and nationally representative household panel datasets: Understanding Society: the UK Household Longitudinal Study (UKHLS) and the China Family Panel Studies (CFPS) (ukhls; cfps).

From a societal perspective, predicting mental health risk can help identify vulnerable groups and support early intervention for people exposed to unstable work situations. From a scientific perspective, testing the added value of unemployment history features helps us understand the mechanisms that connect labor market experiences to mental health, and whether education acts as a buffer or moderator (clark2001unemployment; paul2009does). The study uses a cross-sectional design (one wave from each study) to focus on prediction and comparability. Possible longitudinal extensions will be discussed as future work.

2 literature review

Previous research discovered strong links between unemployment and mental health. For example, Yang et al. (2024) analyzed data from 201 countries between 1970 and 2020 and found that unemployment is significantly associated with higher risks of mental disorders, especially anxiety and depression. These findings indicate that unemployment is not only an economic challenge but also a public health issue.

More detailed studies have investigated how both current unemployment and past unemployment history matter. Using large-scale Finnish register data, Junna et al. (2022) found that current unemployment remains linked to poor mental health even after controlling for stable personal characteristics. They also reported that longer unemployment histories increase risks, especially for men in younger age groups. This highlights the importance of considering both present and past labor market experiences when predicting mental health.

A broader evidence base is provided by systematic reviews and meta-analyses. Sterud et al. (2025) reviewed longitudinal studies and concluded that unemployment increases the risk of mental health problems, while re-employment tends to decrease the risks. Although evidence certainty is sometimes limited, the general pattern supports the idea that transitions in and out of employment shape mental health trend.

In addition, new methodological approaches from data science are increasingly used. Machine learning has been applied to detect depression through social media traces and survey data (Gadzama et al. (2024); Zhang et al. (2025)). These studies demonstrate the potential of advanced algorithms to identify risk factors and predict mental health outcomes, though many focus on social media data rather than labor market dynamics.

In summary, the literature confirms that unemployment is harmful for mental health, that detailed histories matter, and that re-employment can help reduce risks. However, few studies explicitly test the added predictive value of detailed unemployment features compared to a simple unemployment indicator, and even fewer do so across different cultural and labor market contexts. This project aims to fill that gap.

Note: still in lack of literature providing more on methodology.

3 research strategy & research questions

This Thesis addresses the following research question:

To what extent do unemployment features and education level predict individual mental health outcomes in a cross-sectional setting?

Specific sub-questions include:

- Does early-career unemployment have more long-term mental health effects than mid-career unemployment?
- Do detailed unemployment features (duration, frequency, age at first unemployment) improve prediction performance compared to simple unemployment indicators?
- Does grouping the sample by detailed demographic characteristics (e.g., gender, age bands, or urban/rural residence) improve the performance of model prediction?

Hypotheses:

H1: Early-career unemployment is associated with more severe long-term negative effects on mental health than mid-career unemployment.

H2: Higher educational attainment is linked to better mental health outcomes.

H3: Including detailed unemployment features (duration, frequency, age at first unemployment) improves prediction accuracy compared to using a simple unemployment indicator alone.

H4: Grouping individuals by demographic characteristics (e.g., gender, age bands, urban/rural residence) further improves model performance in predicting mental health outcomes.

4 methodology and evaluation

4.1 Dataset Description

This Thesis will use data from two major panel datasets: the China Family Panel Studies (CFPS2022) and Understanding Society (UKHLS). (Not yet access to Finish Dataset. May give up this one)

CFPS. The CFPS dataset(2022) includes 27,001 individual instances. It provides a rich set of variables relevant for this study: six unemployment-related variables, eight employment-related variables, three education-related variables, and nine variables covering mental health and mental illness. This allows detailed modeling of both labor market history and mental health outcomes in the Chinese context.

UKHLS. The UKHLS provides a rich set of employment variables including employment history, hours worked, self-employment, wages, and deductions. Education-related information includes background and attainments, expectations, and recent education and training. For mental health, key variables include hlprbi (self-reported health problems such as anxiety and depression) and hlprxi (treatment for mental health conditions). These measures provide comparable mental health indicators to those in CFPS.

Together, these datasets are large-scale, nationally representative, and public available, making them suitable for comparative and predictive analysis in this Thesis.

Note:Done EDA on CFPS2022, still conduct feature engineering. Preparing EDA on UKHLS.

4.2 Algorithms and Software

The analysis will be mainly conducted using Python. Predictive models will include machine learning methods such as random forests and gradient boosting. Feature engineering techniques will be used to assess the relative contribution of unemployment details versus broad indicators. The use of multiple algorithms ensures robustness and allows comparisons across models.

4.3 Evaluation Method

Model performance will be evaluated using standard predictive accuracy metrics such as RMSE (for continuous outcomes) or accuracy (for binary outcomes). Cross-validation techniques will be employed to avoid overfitting. Comparisons will be made between models with simple unemployment indicators and those including detailed unemployment features, testing whether additional variables significantly improve predictive performance. Grouped analysis (e.g., by gender, age cohort) will also be explored to assess effects of countries with different economy and culture.

5 milestones and plan

- week 1–4: Literature review and dataset familiarization.
- week 4–5: Data preprocessing and feature engineering.
- week 6–9: Model development and evaluation.
- week 10-11: Analysis and check.
- week 12-14: Writing and finalization of the Thesis.

references

Gadzama, W. A., Gabi, D., Argungu, M. S., & Suru, H. U. (2024). The use of machine learning and deep learning models in detecting depression on social media: A systematic literature review. Personalized Medicine in Psychiatry, 45-46, 100125. https://doi.org/10.1016/j.pmip.2024.100125

Junna, L., Moustgaard, H., & Martikainen, P. (2022). Current unemployment, unemployment history, and mental health: A fixed-effects model approach. American Journal of Epidemiology, 191(8), 1459–1469. https://doi.org/10.1093/aje/kwac077

Sterud, T., Lunde, L., Berg, R., et al. (2025). Mental health effects of unemployment and re-employment: A systematic review and meta-analysis of longitudinal studies. Occupational and Environmental Medicine. https://doi.org/10.1136/oemed-2025-110194

Yang, Y., Niu, L., Amin, S., & Yasin, I. (2024). Unemployment and mental health: A global study of unemployments influence on diverse mental disorders. Frontiers in Public Health, 12, 1440403. https://doi.org/10.3389/fpubh.2024.1440403

Zhang, Y., Wang, Z., Ding, Z., Tian, Y., Dai, J., Shen, X., Liu, Y., & Cao, Y. (2025). Employing machine learning and deep learning models for mental illness detection. Computation, 13(8), 186. https://doi.org/10.3390/computation13080186