

Data sheet for paper titled “Silent Suffering: Using Machine Learning to Measure CEO Depression”

Sung-Yuan (Mark) Cheng
University of Kentucky
mark.cheng@uky.edu

Nargess M. Golshan
Indiana University
nargolsh@iu.edu

1. A description of which author(s) handled the data and conducted the analyses.

Nargess Golshan and Mark Cheng both handled the data and conducted different parts of the analyses.

2. A detailed description of how the raw data were obtained or generated, including data sources, the specific date(s) on which data were downloaded or obtained, and the instrument used to generate the data (e.g., for surveys or experiments). We recommend that more than one author is able to vouch for the stated source of the raw data.

Various commercial databases were used for the analyses from the following sources:

- DAIC-WOZ dataset was downloaded from USC Institute for Creative Technologies link received via email on 6/6/2021.
 - Conference call transcripts and audio files for S&P500 firms during 2010-2021 were downloaded from www.capitaliq.com between 7/1/2023 until 9/30/2023.
 - Compustat-CRSP merge annual and quarterly were downloaded from WRDS on 10/17/2023.
 - Execucomp was downloaded from WRDS on 1/27/2024
 - ISS was downloaded from WRDS on 10/7/2023
 - BoardEx was downloaded from WRDS on 9/28/2023
 - IBES Analysts dataset was downloaded from WRDS on 10/17/2023
 - IBES Guidance dataset was downloaded from WRDS on 10/17/2023
- 3. If the data are obtained from an organization on a proprietary basis, the authors should privately provide the editors with contact information for a representative of the organization who can confirm data were obtained by the authors. The editors would not make this information publicly available. The authors should also provide information to the editors about the data sharing agreement with the organization (e.g., nondisclosure agreements, any restrictions imposed by the organization on the authors, such as restrictions to publish certain results).**

For training the depression score estimation model, we used the DAIC-WOZ dataset collected by the USC Institute for Creative Technologies. The dataset is available to academics and other non-profit researchers upon signing an agreement form. We used commercially licensed data available through WRDS and the CapitalIQ Browser Platform for the rest of the analyses.

4. **A complete description of the steps necessary to collect and process the data used in the final analyses reported in the paper. For experimental and survey papers, we require information about the instructions and instruments used to generate the data, subject eligibility and/or selection, as well as any exclusion criteria. The full set of instructions and instruments can be provided in the online appendix.**

The process to collect, process, and merge the datasets together and run the analyses is as follows:

- Download DAIC-WOZ dataset.
 - Train machine learning model to estimate depression score.
 - Download audio files and transcripts of conference calls from CapitalIQ Browser Platform.
 - Diarization of audio files to extract the part that CEO speaks during the presentation part.
 - Process the CEO part audio files and deploy the trained model to estimate CEO Depression Score.
 - Collect data from sources mentioned above and detailed in the main text of the paper to create the variables needed for the analyses.
 - Merge datasets together.
 - Run the regression models.
5. **The computer programs or code used to convert the raw data into the final dataset used in the analysis plus a brief description that enables other researchers to use this program. The purpose of this requirement is to facilitate replication and to help other researchers understand in detail how the raw data were processed, the final sample was formed, variables were defined, outliers were treated, etc. This code or programming is in most circumstances not proprietary. However, we recognize that some parts of the code or data generation process may be proprietary, including from the authors' perspective. Therefore, instead of the code or program, researchers can provide a detailed step-by-step description of the code or the relevant parts of the code such that it enables other researchers to arrive at the same final dataset used in the analysis. In such cases, the authors should inform the editors upon initial submission, so that the editors can consider an exemption from the code sharing requirement. Whenever feasible, authors should**

also provide the identifiers (e.g., CIK, CUSIP) for their final sample. Authors should consult our FAQ Sheet on the JAR website for further details.

We share the codes and a README file that details the order for running them online.

6. An assurance that the data and programs will be maintained by at least one author (usually the corresponding author) for at least six years, consistent with National Science Foundation guidelines.

We agree to maintain the data and programs for six years.