MENTAL HEALTH IN TECH

Group D9: Karel Allik, Hendrik Jaks, Robin Juul

Repository link

Business understanding

Background

Since tech industries continue to shape the future of work and society, it is important to take note of the mental health of the people working in the industry. It is widely known that working in tech often involves long hours, high cognitive loads, and remote work. Thus, it is critical to understand the well-being of the workforce. These factors can contribute to burnout, anxiety, and the development of mental health disorders. By analyzing this data and expanding it, we can provide insight into mental health trends and additionally identify the factors that strongly correlate with mental health disorders.

Business goals:

- Raise awareness of mental health in the tech industry.
- Help identify risk factors and reduce stigma surrounding mental health.
- Track long-term trends in mental health in the tech industry.

Success criteria:

- Project produces a report and visualization that highlights the mental health trends.
- Analysis reveals specific factors that correlate with mental health outcomes.
- Clear year-over-year comparisons of mental health outcomes and workplace attitudes.

Inventory of Resources

- People: Team members (Karel Allik, Hendrik Jaks, Robin Juul).
- Data: From Open Source Mental Illness (OSMI) using survey data from years 2014, 2016, 2017, 2018 and 2019 with the expansion of the recent years 2020 – 2023. Each survey measures attitudes towards mental health and frequency of mental health disorders in the tech workplace.
- Hardware: Access to modern computing resources with sufficient memory and processing power to handle large-scale text preprocessing and model training.

- Software: Python ecosystem (pandas, NumPy, scikit-learn), Hugging Face Transformers for NLP models (Sentence-BERT, RoBERTa), SQLite for database management, and visualization libraries (matplotlib, seaborn).

Requirements:

- Schedule for completion: The project must be finished by the 8th of December.
- Deliverables: a poster and a project poster PDF file.
- Data access: the team must be able to use the OSMI survey data from 2014–2019 and 2020–2023.
- Tools: work must be done with approved and licensed software (Python, SQLite, Jupyter).

Assumptions:

- The OSMI survey data represents a broad sample of the tech industry to draw meaningful insights.
- The additional 2020–2023 responses are consistent in format with earlier surveys.
- Team members have sufficient technical skills to preprocess text and apply machine learning models for classification and prediction tasks.

Constraints:

- The surveys from recent years have far less data, which may limit the statistical power of our analysis.
- Time is also of the essence, as the project must be completed within a fixed schedule, leaving limited room for experimentation.
- Computational power could become an issue since the dataset contains over 240 000 rows of which most are open-ended text responses. Analyzing and working with such data requires a lot of memory and resources, especially when using LLM models.

Risks:

- Low response rates for recent surveys can limit the representativeness of the dataset.
- Data quality issues like missing values, inconsistent formatting, or too vague answers may reduce the reliability of the analysis.

Terminology:

- Mental health disorder. A condition that affects a person's emotional, psychological, or behavioral well-being (e.g., anxiety, depression).
- Stigma. Negative beliefs, attitudes, or stereotypes about mental health issues.
- Risk factors. Workplace or personal conditions that increase the likelihood of developing a mental health disorder (such as high cognitive load, long working hours, or limited support in remote work).
- Feature. A measurable characteristic used for analysis or prediction.
- Label. The outcome the model is trained to predict.
- Classification model. A machine learning model that categorizes data into predefined classes.
- Sentence-BERT / RoBERTa. Transformer-based NLP models used to embed text and classify language.
- Precision. Measures how many predicted positives are correct.
- Recall. Measures how many actual positives are successfully detected.

Costs:

- We plan to execute the project using only our personal computers, which eliminates hardware and infrastructure costs.
- If additional compute power is needed, we may use the free Google Gemini API to support text analysis tasks.

Data-mining goals:

- Expand the current dataset with additional data from the years 2020-2023.
- Predict whether a respondent has a mental health disorder (Random Forest / Deep Learning with Sentence-BERT).
- Build a classifier that detects stigmatizing vs supportive language in text answers (RoBERTa) and analyse the result on a yearly basis.

Data-mining success criteria:

- Ensure the new data is consistent with the earlier surveys.

- Insights reveal which features are most influential and achieve strong predictive performance (accuracy ≥ 80% with balanced precision and recall).
- Yearly analysis shows clear trends in language use, with precision and recall ≥ 90% for both classes.

Data understanding

Data requirements

- Quantitative variables (frequency of mental health disorders, demographics, company size).
- Qualitative variables (open-ended text responses about experiences, stigma, and support).
- Surveys from 2014–2019 (OSMI) plus extended responses from 2020–2023 to enable longitudinal analysis.
- Structured tables (CSV, SQLite) for numeric and categorical variables.
- Text fields for open-ended responses, suitable for NLP preprocessing.
- Must include identifiers (survey ID, question ID, user ID) to link across tables.

The data is sourced from Kaggle, which is originally from Open Source Mental Illness (OSMI) surveys conducted in 2014, 2016, 2017, 2018, and 2019. Additional responses from 2020–2023 are being collected to extend the dataset. All data is available on Kaggle or on OSMI official page.

Selection criteria:

Data sources:

- Database file including surveys from 2014- 2019
- Four .csv survey files from years 2020-2023 (all data is initially used and later filtered depending on the situation and suitability)

Database tables and fields used:

- Survey table — survey ID and description, covering surveys from 2014, 2016, 2017, 2018, and 2019.
- Question table — question ID and text, containing 105 unique survey questions.
- Answers table — respondent ID, survey ID, question ID, and answer text, with ~250,000 rows.

Case ranges:

- All respondents across survey years 2014–2019.
- Additional responses collected from 2020–2023 to enable year-over-year comparisons (from OSMI page).

- Include both individuals with diagnosed mental health disorders and those without, as data from both groups is essential for predictive modeling.

Describing data

The data from the years 2014-2019 has approximately 240 000 rows of data representing answers from around 4200 individuals. Additionally, we have surveys from 2020 to 2023 and those add about 500 more participants. The surveys from 2014-2019 have a total of 105 different questions, of which many are open-ended, resulting in many open-ended answers. Over the years the surveys don't have all the same questions and therefore, we must decide what to do with them. However, since we have the following question in the survey: "*Do you currently have a mental health disorder?*", we are able to use it as a label for predicting mental health disorders and since we also have a lot of open-ended answers we can use them to recognize stigmatizing or supportive language in the answers.

Exploring data

Looking at the data from the years 2014-2019, we can see that there are only about 3000 answers from the participants to the question "*Do you currently have a mental health disorder?*". This was probably caused by the fact that in the 2014 survey the question was most likely not asked. Therefore, we might have to exclude the answers from 2014, since we are not able to use this data to achieve the goal of predicting whether a respondent has a mental health disorder or not. Since there are many open-ended questions, the answers should be combined into single text, so that later the models would not have to look at every answer separately but rather see the bigger picture.

The data has also several quality challenges, that must be addressed before analysis. Combining all the questions and answers from the years 2014-2019 into a single csv file, we see that many cells are NaN, because the questions vary over the years and there are many cells containing –1, indicating missing values. Looking at the gender column there are many different answers indicating male, for example there are values such as "male", "cishet male", "male-ish", "ostensibly male" etc, which all must be cleaned.

Also, the surveys from 2020 to 2023 contain similar, but not exactly the same questions as the ones in 2014-2019, which must be addressed before starting to use the data. Looking at the data from recent years (2020-2023) there also appears to be many NaN's and inconsistencies in the participants' answers, meaning that some individuals did not answer all of the questions.

Verifying data quality

Despite quality challenges, there is enough data, a total of about 2900 different individuals, that have answered the question whether they have a health disorder or not. From all of the questions in the surveys, we are able to limit them down to the most answered and impactful ones and use the filtered dataset for the classification task. Regarding the NLP-based analysis we can use the open-ended text from the participants to successfully work towards and achieve this task. Looking at the goal of expanding the dataset with responses from 2020–2023, the main challenge lies in the variation and differences of survey questions across years. However, this issue is doable given the large number of questions overall, allowing for alignment during the preparation.

In conclusion, the previously defined data-mining goals can be successfully reached once the data quality is resolved and the dataset has been prepared for analysis.

Plan for the project

- Project idea generation and planning – Karel & Hendrik 3h
- Report (HW 10) - Karel 6h, Hendrik 2h

Acquire and expand data – Hendrik & Robin 6h each, Karel 4h

- Download the dataset from Kaggle
- Inspect columns, data types, missing values
- Collect additional datasets for the years 2020–2023
- Align and merge datasets into a unified format

Explore dataset – Hendrik & Karel 2h each

- Inspect combined dataset
- Identify the target variable for "mental health disorder" prediction
- Identify text fields relevant for language classification

Clean and reprocess data Hendrik, Karel & Robin 5h each

- Handle missing or inconsistent values
- Clean and standardize categorical fields (gender, company size, etc.)
- Normalize or scale numerical features where needed
- Clean and prepare text fields (remove noise, fix encoding issues)
- Ensure consistency across older + newer datasets

Feature engineering – Karel 3h

- Encode categorical variables (One-Hot / Ordinal)
- Select relevant numerical features
- Identify and prepare text fields for NLP analysis
- Split dataset into train, validation, and test sets
- Select relevant features for mental-disorder prediction

Text embedding with Sentence-BERT – Hendrik 3h

- Load a pre-trained Sentence-BERT model

- Generate embeddings for selected text fields
- Integrate embeddings into your feature set

Random Forest model – Karel 3h

- Train a Random Forest classifier on structured data
- Perform hyperparameter tuning
- Evaluate performance on validation data

Deep learning model – Hendrik 4h

- Build a classifier using SBERT embeddings
- Train, tune, and evaluate the model
- Evaluate and compare performance with Random Forest

Stigmatizing vs supportive language classification (RoBERTa) - Robin 6h

- Define how to label text as stigmatizing or supportive.
- Create a dataset of stigmatizing vs supportive text
- Fine-tune a RoBERTa text classification model
- Train, validate, and test the classifier
- Analyze which phrases or patterns differentiate stigmatizing vs supportive responses

Yearly analysis of stigma trends (using RoBERTa output) - Robin 6h

- Apply the trained RoBERTa model to all years
- Compute yearly proportions of:
  - stigmatizing language
  - supportive language
- Visualize trends across years

Model evaluation and comparison – Hendrik, Karel 3h each

- Compute metrics: accuracy, precision, recall, F1, ROC-AUC
- Compare Random Forest vs deep learning model performance

- Perform error analysis

Model explainability and interpretation –Robin, Hendrik 3h each

- Check which features matter most in the trained models
- See which words in a sentence influence RoBERTa's decision.
- Interpret which words indicate stigma or support.

Poster preparation, submission, and project presentation – Hendrik, Robin & Karel 4h each

- Create the final project poster (A0 portrait or A1 landscape) summarizing:
  - Dataset expansion process (2014–2023)
  - Stigma trends over the years
  - Model results and comparisons
  - Key visualizations (confusion matrices, ROC curves, yearly trend plots)
  - Main conclusions and insights
- Prepare for the poster session presentation

Tools and methods

- Sentence-BERT – used to generate text embeddings that capture sentence meaning for classification tasks.
- RoBERTa – a transformer-based model to classify stigmatizing vs supportive language.
- Python, pandas, NumPy – for reading, cleaning, and processing data
- matplotlib.pyplot – for creating graphs and visualizations
- scikit-learn – used for encoding categorical values, splitting data into training/testing sets, building a prediction model, and evaluating model performance
- glob – for loading multiple files automatically (e.g., datasets from different years)
- SequenceMatcher (difflib) – for detecting and fixing similar or inconsistent text values
- sqlite3 – to access data in a database
- Google Gemini API – used for automated text classification and labeling