Group Members: Anita Gee, Siwei Guo, Yingzhu Chen, Nemo (Sorachat) Chavalvechakul
Group Name: Mental Health Warriors

Problem Statement

The increasing use of social media and online interactions today has a significant impact on mental health, often reflected through textual posts that may indicate various mental health conditions. It's crucial to develop analytical tools capable of detecting and categorizing these indicators accurately. This project focuses on creating solutions that can identify and classify mental health statuses from text data, helping to improve the effectiveness of mental health chatbots and providing insights into broader mental health trends.

Articulation of value - Ying

According to the WHO, "*mental health is a state of mental well-being that enables people to cope with the stresses of life, realize their abilities, learn well and work well, and contribute to their community.*"

Mental health conditions include mental disorders and psychosocial disabilities as well as other mental states associated with significant distress or even risk of self-harm.

According to WHO data, "*In 2019, 970 million people globally were living with a mental disorder, with anxiety and depression the most common. Globally, mental disorders account for 1 in 6 years lived with disability. People with severe mental health conditions die 10 to 20 years earlier than the general population. And having a mental health condition increases the risk of suicide and experiencing human rights violations.*"

Mental health disease can cause difficulty in many aspects of life, including relationships with family, friends, and communities. This will lead to possible problems in the school or working environment. The economic loss that mental health disease causes is enormous. From the article "The Global Economic Burden of Non-communicable Diseases", Bloom etc. estimated that the economic loss caused by mental health disease in the year 2021 was roughly 1·3 trillion US dollars (USD) in 2010 ($1·6 trillion in 2019). The author predicts that in 2030, the losses would grow to around 2·5 trillion USD in 2010 (or approximately 3 trillion USD in 2019). In the article "Quantifying the global burden of mental disorders and their economic value", the author also mentioned that the health and economic burden caused by mental health issues might be even much higher than previously estimated.

With NLP techniques, it gives us the chance to recognize potential mental health issues from their social media posts. We can lower the health and economic burden through earlier early intervention. By identifying individuals at risk and early intervention, sentiment analysis can help reduce the burden on the public health system, future optimal resource allocation, reduce healthcare costs, and contribute to positive economic outcomes.

References:

https://www.thelancet.com/journals/eclinm/article/PIIS2589-5370(22)00405-9/fulltext

chrome-extension://oemmndcbldboiebfnladdacbdfmadadm/https://www3.weforum.org/docs/WEF_Harvard_HE_GlobalEconomicBurdenNonCommunicableDiseases_2011.pdf

https://www.who.int/health-topics/mental-health#tab=tab_2

Calculation of the potential economic value - Ying

Calculation idea for now:
1: get the average mental health illness medical cost for each patient: Data source WHO?
2: identify the productivity loss caused by mental health issues - need research support (haven't finished reading document yet)
3: the percentage of successfully identified people who may have mental health issues - from the model/ dataset

4: How would early intervention affect mental health issues? - The percentage of people who get early intervention vs people who did not get early intervention?

Calculation for medical cost savings:
Medical cost saving = targeting people amount * % of early intervention * average medical cost of mental health disease * % of effetion from early intervention.
People amount * index 1 * percentage 3 * percentage 4

The study, *The Effect of Messaging Therapy for Depression and Anxiety on Employee Productivity*, highlighted that the average cost of traditional therapy is $1,120 for a 3.5-month treatment period. Projecting this over a 12-month period, the annual cost per patient for traditional therapy would amount to:

1,120 USD × 12 / 3.5 = 3840 USD / year

This figure assumes consistent therapy engagement throughout the year at the same rate.

Assume the accuracy of our model is 80%, percentage of successful early intervention is 30%

Accordion to the data from https://www.nimh.nih.gov/health/statistics/mental-illness

, “It is estimated that more than one in five U.S. adults live with a mental illness (59.3 million in 2022; 23.1% of the U.S. adult population).”, our people amount will be 59.3 million.

Our maximum estimated total productivity loss will be
59.3 million * 80% * 30% * $3840 = $54,650,880,000. (around 54 billion)

Reference:

Talkspace Annual Report:
https://investors.talkspace.com/static-files/90d8b389-26cb-40be-81de-01ff24e8f72e

The Effect of Messaging Therapy for Depression and Anxiety on Employee Productivity

Journal of Technology in Behavioral Science (2019) 4:1–5

Calculation for Productivity loss decreasing
Productivity loss decreasing = targeting people amount * % of early intervention * average productivity decreasing of mental health disease * % of effetion from early intervention.

People amount * index 2* percentage 3 * percentage 4

https://www.gallup.com/workplace/404174/economic-cost-poor-employee-mental-health.aspx

The article mentioned that “Projected over a 12-month period, workers with fair or poor mental health are estimated to have nearly 12 days of unplanned absences annually compared with 2.5 days for all other workers. Generalized across the U.S. workforce, this missed work is estimated to cost the economy $47.6 billion annually in lost productivity.”

There’s a 12 - 2.5 = 9.5 days difference between the employee have mental health issue compare to regular employees, we can use average wage/GDP to identify the productivity loss.

For example, assume the average wage is $25, total loss for a person a year would be $25/h *8 hours * 9.5days = $1900

Assume the accuracy of our model is 80%, percentage of successful early intervention is 30%

Accordion to the data from https://www.nimh.nih.gov/health/statistics/mental-illness

, “It is estimated that more than one in five U.S. adults live with a mental illness (59.3 million in 2022; 23.1% of the U.S. adult population).”, our people amount will be 59.3 million.

Our maximum estimated total productivity loss will be
59.3 million * 80% * 30% * $1900 = $27,040,800,000 (around 27 billion)

<u>Final economic impact = medical cost saving + productivity loss decreasing</u>

Project Plan

Week	Steps	What we will do
Week 1	Ingest and Explore the Dataset	We will identify the dependent variable (predictor). In our dataset, this will be the status column.We will load the data into JupyterHub and look at the number of rows and columns, see the data types, and see if there are any missing values. If there are missing values, come up with an imputation method to fill those in (mean, median, backfill or forward fill etc.). Check for any duplicates. Generate summary statistics. Create a written report, Jupyter Notebook with code and output of code. Merge report and code into Github Repository.
Week 2	Perform Exploratory Data Analysis	We will conduct exploratory data analysis using Python in Jupyter Hub. For exploratory data analysis, we will provide visualizations of the distribution of different mental health statuses to understand class imbalances. Create word clouds for each mental health status to visualize the most frequent words. Analyze the distribution of text lengths to understand the variability in statement lengths. We will then split the dataset into training, validation and test datasets. Identify the preprocessing steps. Create a written report, Jupyter Notebook with code and output of code. Merge report and code into Github Repository.
Week 3	Make Data Model Ready Preprocessing	We will preprocess the data using our text normalization function. First, we will initialize a tokenizer (i.e TokTok) and a list of English stopwords, strip HTML, apply stemming (reducing words to its base or root form), remove repeated characters, accented characters and special characters, expand contractions and remove stop words. Next, we will initialize a vectorizer (i.e. Bag of Words, TF-IDF, or transformer models such as BERT) to turn the text data into numerical data. Create a written report, Jupyter Notebook with code and output

		of code. Merge report and code into Github Repository.
Week 4	Engineer Features	We will create and engineer new features by creating a new cleaned text column and deleting the uncleaned text column. We will one-hot encode the status column to make it easier to create machine learning models. We will also create a text length column. Create final training, validation and test datasets that will be used in modeling, evaluation and testing. Create a written report, Jupyter Notebook with code and output of code. Merge report and code into Github.
Week 5	Develop a 1st modeling approach (simple, the baseline)	Build a very simple set of models appropriate for sentiment analysis of mental health such as multinomial logistic classification problem, SVM model, and Naive Bayes. Tune the hyperparameters for these models to get the highest accuracy. We select model evaluation metrics such as accuracy and f1 score. We evaluate variations based on accuracy and f1 score as well as the confusion matrix and pick the best model with the highest f1 score and accuracy on the validation set. Create a written report, Jupyter Notebook with code and output of code. Merge report and code into Github Repository.
Week 6	Build a more complex set of models. Develop a 2nd modeling approach.	Develop more complex models and finetune the hyperparameters of all models for the best accuracy. We will develop models such as Random Forest, K-Nearest Neighbors (KNN) and Gradient Boosting Machines (GBM) such as XBGoost. We select model evaluation metrics such as accuracy and f1 score. We evaluate variations based on accuracy and f1 score as well as the confusion matrix and pick the best model with the highest f1 score and accuracy on the validation set. Create a written report, Jupyter Notebook with code and output of code. Merge report and code into Github Repository.
Week 7	Develop the most complex set of models. Develop a 3rd modeling approach.	Develop the most complex models and finetune the hyperparameters for all models. We will use LSTM models, RNN models, transformer models such as BERT, DRoberta and DistilBERT. We select model evaluation metrics such as accuracy and f1 score. We evaluate variations based on accuracy and f1 score as well as the confusion

		matrix and pick the best model with the highest f1 score and accuracy on the validation set. Create a written report, Jupyter Notebook with code and output of code. Merge report and code into Github.
Week 8	Select the Winning Model.	Evaluate multiple models. Select the best model based on highest accuracy score and f1 score. Predict on the test set. Make sure that the test dataset is preprocessed in the same manner as the train and validation dataset. Calculate performance on the test dataset such as the accuracy and f-1 score on the test set. Create a written report, Jupyter Notebook with code and output of code. Merge report and code into Github Repository.
Week 9	Data Centric AI	Improve your model by improving the data. Consider more preprocessing steps such as lemmatization to improve performance or remove preprocessing steps to see whether less cleaned data may improve performance. Consider preprocessing emojis and emoticons. Consider augmentation techniques such as synonym replacement and shuffling sentences. Consider finding another dataset to augment the dataset. Create a written report, Jupyter Notebook with code and output of code. Merge report and code into Github Repository.
Week 10	Explain the model, analyze risk, bias and ethical considerations	Explain your model by understanding feature importance and prediction outcomes. We will use the random forest model to identify some important features and the confusion matrix to identify false positives and false negatives. Identify model risks such as overfitting by monitoring epochs and performing cross validation. If the validation loss and validation accuracy does not improve with each epoch but the training accuracy and loss continues to decrease, it indicates overfitting. We will put in early stopping parameters to prevent this. Identify and quantify bias in your input dataset and model output. We will need to make sure that the labeling process is unbiased and consistent. Identify and measure bias by checking whether the dataset has diverse representation. Ethical considerations of diagnosing someone’s mental health status could be damaging to

		both providers, patients and their family and friends. We must be careful to not have false positives and false negatives. Create a written report, Jupyter Notebook with code and output of code. Merge report and code into Github Repository.
Week 11	Save and package your model for deployment. Build your model monitoring plan.	Use pickle or joblib packages for traditional machine learning models. Use torch.save for Pytorch and transformer models. Create a requirements file to list all dependencies. Package my model into a streamlit app. For model monitoring, set up logging, tensorboard or MLFlow and monitor data drift. Lastly, set up alerts if accuracy drops below a certain threshold. Early Stopping parameters will prevent overfitting. Create a written report, Jupyter Notebook with code and output of code. Merge report and code into Github Repository.
Week 12	Put it all together!	Finalize the full project by cleaning up any errors in code and making sure that the report and code are easy to read and follow. Make a final decision on the best model. Create a written report, Jupyter Notebook with code and output of code. Merge report and code into Github.
Week 13	Peer Review	Review a peer's work and related artifacts. Provide written feedback reports to a peer's work and artifacts.

Discuss the Dataset

For this project, we have chosen the "Sentiment Analysis for Mental Health" dataset available on Kaggle. This dataset is a well-curated collection from various sources, designed specifically for training machine learning models to perform sentiment analysis related to mental health.

Dataset Description:

- **Source**: Compiled from multiple Kaggle datasets, this dataset includes data from discussions about depression, stress prediction, anxiety, and other mental health issues, gathered from platforms like Reddit and Twitter.
- **Content**: The dataset consists of textual entries, each labeled with a mental health status such as Normal, Depression, Anxiety, etc. This labeling supports the development of models aimed at accurately classifying these conditions.

- **Utility**: This dataset is particularly useful for developing sophisticated mental health chatbots and conducting in-depth sentiment analysis. The diversity of the data ensures robust training for models, enabling them to recognize and predict a range of mental health conditions from textual inputs.

Relevance to the Problem Statement:

This dataset perfectly matches our project's goals as it provides extensive data necessary for training sentiment analysis models. By leveraging this dataset, we can enhance the algorithms that detect and categorize mental health statuses from texts, which is pivotal for augmenting mental health chatbots and gaining deeper insights into mental health patterns. This will ultimately aid in more effective mental health interventions.

Reference:

https://www.kaggle.com/datasets/suchintikasarkar/sentiment-analysis-for-mental-health

Discuss Modeling Techniques

For sentiment analysis on mental health here is the appropriate modeling approach:

- Type of Modeling: Supervised Learning
- Classification or Regression: Classification
- Binary or Multi-Class: Multi-Class Classification

Because the data consists of statements tagged with one of seven mental health statuses (Normal, Depression, Suicidal, Anxiety, Stress, Bi-Polar, Personality Disorder), this makes it a supervised learning problem, as the data is labeled. Second, the goal is to classify each statement into one of the seven categories, making it a multi-class classification problem.

We will use common NLP (Natural Language Processing) techniques such as the nltk package in Python. We will also preprocess the text by creating a normalization function, which applies a series of text normalization steps. First, we will initialize a tokenizer (i.e TokTok) and a list of English stopwords, strip HTML, apply stemming (reducing words to its base or root form), remove repeated characters, accented characters and special characters, expand contractions and remove stop words. Next, we will initialize a vectorizer (i.e. Bag of Words, TF-IDF, or transformer models such as BERT) to turn the text data into numerical data. By using train-validation-test splits, we will then fit our models, evaluate its accuracy to predict sentiment analysis and then test on an unseen test dataset.