Prediction of Mental Health with Natural Language Processing

Proposal

Submitted by:

Sepideh Azadian

Ayse Nur Safak

Seyedeh Shadi Sadeghinik

Yidi Zhao

MSc Data Science
[University of Europe for Applied Sciences]

Table of Contents

Abstract

Anxiety disorder is one of the most common health problems today, but still, a lot of people try to diagnose their problem early. Some factors, like a lack of awareness and limited access to specialists, prevent early help. This project suggests a different way of diagnosing this problem by using machine learning to analyse people's own words. By analysing words that contain individuals' feelings and daily experiences, our system wants to predict the signs of mental problems automatically.

In this study, we use hybrid models that combine BERT with XGBoost for classification. BERT is used to understand the deep meaning and emotional information from the text, and XGBoost is used for classification. In this way, we can prepare a process with high accuracy and easy to interpret for predicting the mental problem from self-reported statements. We used over 53000 self-reported statements to make a reliable and interpretable system that can help with the early detection of mental health problems. Our final goal is to create a scalable system that can support mental health problems and help individuals seek help sooner.

Introduction and Background

Mental health is the most widespread health challenge nowadays. It affected all people of all ages and with all backgrounds. Despite the awareness of people, mental health problems often remain undetected and untreated, which causes a serious impact on an individual's well-being and quality of life. Traditional diagnosis of mental health contains personal interviews, standardised questionnaires, and clinical observations. While these methods are effective, they are costly and time-consuming and require access to experts. In addition, many people avoid seeking these helps due to their stigma or unawareness, which makes early detection impossible.

Recently, according to good advances in artificial intelligence and especially in machine learning (ML) and natural language processing (NLP), there are different ways to understand mental health problems according to language. People often express their feelings, emotions and thoughts in social media or anonymous surveys. By analysing these texts, which contain a lot of signs and analysing them, we can produce a scalable and cost-effective way to detect mental problems. This project builds on this idea by using hybrid models that are a combination of BERT with XGBoost. BERT helps us

understand the meaning in mental problem-related text, and XGBoost perform reliable predictions.

Problem Statement

These days, mental problems affect millions of people around the world and their quality of life. Traditional ways of diagnosing, like a questionnaire or a clinical interview, are now accessible, but they are not always practical and accessible to everyone. Some people even avoid this expert assessment because of the fear of judgment, or maybe they cannot see how serious their problem is. Sometimes, early symptoms of mental problems are hard to diagnose, and a delay in treatment will cause irreparable harm in people's lives. So, there is an urgent need for a solution that can help identify even unnoticed symptoms very fast.

Besides all progress in machine learning and NLP, there are still some gaps that this thesis aims to cover:

1. Limited focus on one special mental problem: We want to cover more mental problems with this self-reported dataset.
2. Not enough studies using deep meaning understanding from advanced models with powerful classifiers on the rich mental health dataset.
3. Insufficient work on explainability makes that challenging to understand the model prediction in important cases.
4. Limited work about the reliability of models for different groups and text types.

By mentioning these gaps, our project develops interpretable and accurate machine learning models for detecting mental problems with an early and interpretable approach.

Research Objectives

This research will pursue the following specific objectives:

1. Create an accurate and scalable machine learning model
2. Using a hybrid model that combines BERT for deep meaning representation and XGBoost for classification.
3. Design a text processing pipeline which preprocesses the text, punctuation, stop words, and stems to improve the model efficiency.
4. Combine different techniques to clarify how models make predictions and make sure the system is interpretable.
5. To evaluate the text is generalised to different user groups
6. Prepare an interpretable system which can help the early mental problems detection

Theoretical Framework

This research is formed at the intersection of two important fields: cognitive science and clinical psychology on the one hand, and machine learning on the other hand. The theoretical framework of the project is based on the following concepts:

1. Psychological theories:

Considering cognitive psychology research, the mental and emotional states of any individual can be a reflection of their language and expressions they use in their daily life. Those who are afflicted with cognitive disorders such as anxiety, depression, or OCD usually use specific kinds of words, verbs, or sentence structures. With this in mind, natural language analysis (NLP) can be used as a non-invasive tool for the initial and automated screening of mental disorders.

Also, it is obvious from modern psychoanalytic theories that vocabulary choice, the degree of self-referentiality, negative/positive words, and other linguistic elements are related to mental health status.

2. Machine learning theory:

This project uses a supervised learning model, as the model should learn from labelled data. The BERT + XGBoost hybrid model has shown awe-inspiring performance in classifying psychological texts. This model first extracts semantic and deep features of texts using BERT, one of the most powerful natural language processing models. Afterwards, thanks to the XGBoost model, which is a powerful boosting algorithm in machine learning, these features are used to predict the final classification.

3. Relationship of theories:

The combination of these two perspectives has led to the creation of a new approach in psychological analysis; That is, using machine learning algorithms to discover hidden mental patterns through language. Therefore, this project is based on the key idea that:

"Linguistic patterns can be modelled, and psychological cues in texts can be detected through machine learning."

The theoretical framework of this project, based on the connection between language and mind in psychology and the ability of intelligent models to understand and classify text patterns, provides a platform for effective and automated analysis of mental disorders through written data.

Methodology

6.1 Data Sources

In this study, we plan to explore how people express their mental health experiences through language by using a large dataset, which is called Sentiment Analysis for Mental Health (Sarkar, 2022), was taken from Kaggle. This collection was gathered from personal statements from a wide range of sources, like Reddit posts, Twitter updates, student surveys, and online conversations. This integrated information is from the following Kaggle datasets:

- 3k Conversations Dataset for Chatbot (Rajani, 3K Conversations Dataset for ChatBot [Dataset], 2023)
- Depression Reddit Cleaned (InfamousCoder, n.d.)
- Human Stress Prediction (Rajani, 2023)
- Predicting Anxiety in Mental Health Data (Patricia, 2024)
- Mental Health Dataset Bipolar (Michelle Velice Patricia, 2024)
- Reddit Mental Health Data (Neel Ghoshal, 2023)
- Students Anxiety and Depression Dataset (Sourav Saha, 2022)
- Suicidal Mental Health Dataset (Aradhak Kandhari, 2024)
- Suicidal Tweet Detection Dataset (Syeda Aunanya Mahmud, 2023)

Each statement includes textual statements as sentences, tagged with one of seven specific mental health statuses:

  - Normal
  - Depression
  - Anxiety
  - Stress
  - Bi-Polar
  - Personality Disorder

This dataset itself is not a lone source but an amalgamation of several smaller collections that have been cleaned and combined with richer, more representative resources. It includes contributions from datasets focusing on depression discussion, stress prediction, suicidal ideation, and other mental health topics. Because it captures real language from social media and other online spaces, it offers a window into how people naturally talk about their feelings and challenges.

Each entry in the data has a unique identifier for each entry, the raw text of statements, which was composed of sentences, and the assigned mental status as a classification category.

Equally important will be thinking about the ethical aspects of this work. Even though the data is anonymised and drawn from public sources, it reflects deeply individual experiences. We want to approach with respect, making sure the models aren't used to make inappropriate or oversimplified claims about mental health. We will open about the limitations of any automated analysis, recognising that no model can fully capture the complexity of human emotions or replace professional assessments.

6.2 Preprocessing Pipeline

Our raw data gathered from social media is an informal and often noisy nature of online text, with inconsistent casing, punctuation, informal spelling, and domain-specific slang. We design a robust preprocessing pipeline to clean and standardise the data while preserving psychologically meaningful content.

6.2.1 Handling Missing Values

Working with text data at scale, especially when it is collected from multiple heterogeneous sources, inevitably leads to gaps or inconsistencies. Sometimes entire text entries were blank; in other cases, labels indicating the health status of the post were missing or corrupted. These issues can arise from the original data collection process, errors in manual labelling, or data integration during the compilation of multiple sub-datasets into a single resource.

To address this, we examined the dataset systematically for any null or missing values. We found 362 statements as null. We have 52681 entries. The 362 null data points do not set importance for our studies. So, we dropped the null entries from our dataset. We removed any record lacking the actual text of the statement, as there is no meaningful input for a text classification task without it.

6.2.2 Lowercasing

We begin with text preprocessing with normalisation, converting all text to lowercase to avoid artificial distinctions between capitalised and uncapitalised words. This simple step reduces sparsity in the representation without any loss of semantic content, ensuring "Anxiety" and "anxiety" are treated identically.

6.2.3 Punctuation and Non-Alphabetic Removal

We remove the punctuation and non-alphabetic characters, which are Unicode Characters, Unwanted Patterns and URLs. While these characters often introduce noise, care is taken to ensure that emotive punctuation, such as repeated exclamation marks, is not systematically stripped of such signals, as this could prove useful in later models.

6.2.4 Tokenisation

We utilised tokenisation to segment each statement into individual words. Using nltk.word_tokenize (), we split text into individual tokens, enabling further manipulation at the word level.

6.2.5 Stopwords Removal

We remove stopwords, which are commonly used words that generally offer little discriminative value (e.g., ‘are’, ‘is’, ‘the’, etc.). However, we select the stopword list carefully, recognising that in mental health language, certain stopwords might carry important negations or emotional cues (e.g., ‘not’, ‘non’, ‘not happy’). We avoid removing these words that may carry emotional valence.

6.2.6 Stemming

We apply stemming to reduce words to their root forms, consolidating variations of the same term. For instance, ‘running’, ‘runs’ to ‘run’. This step reduces the overall dimensionality of the feature space while maintaining the core semantic content of the statements. We used the NLTK’s PorterStemmer for this process.

6.2.7 Part-of-Speech (POS) Analysis

To deepen the linguistic analysis beyond simple word frequencies, this study actively applies Part-of-Speech (POS) tagging to the cleaned and tokenised text using the Natural Language Toolkit’s pos_tag function, which assigns grammatical labels such as nouns, verbs, adjectives, and adverbs to each token in the dataset, using established tools designed for English language processing. We analysed each post to generate a sequence of grammatical tags, which we can then examine both qualitatively and quantitatively. By performing POS tagging on preprocessed text, we capture structural patterns in language that may signal distinct mental health statuses. For instance, the model can find differences in the prevalence of personal pronouns, which often indicate self-focus associated with depression, or frequent use of modal verbs suggesting anxiety and uncertainty. We used this grammatical information for exploratory analysis, comparing distributions of POS categories across classes to reveal stylistic differences. By systematically incorporating POS-tag-derived features, the study uses subtle ways in which people construct language when describing mental health experiences, supporting a more nuanced and effective classification model.

6.3 Model Design and Feature Engineering: BERT+XGBoost

To truly understand how people talk about their mental health, this study moves beyond simple keyword counting to a more advanced approach that combines deep language understanding with flexible, interpretable classification. We achieve this through a pipeline that uses BERT (Bidirectional Encoder Representation from Transformers) and XGBoost (Extreme Gradient Boosting) for classification and prediction of mental status,

chosen specifically to capture the subtlety and complexity of mental health language in everyday use.

BERT (Devlin, 2019) is a model trained on vast amounts of text to learn how words mean different things depending on context. Unlike similar methods that treat words in isolation, BERT reads an entire sentence at once, considering the words before and after to understand nuanced meaning. This is important in mental health contexts where meaning can be subtle, indirect, or emotionally charged. For instance, the phrase "I can't do this anymore" carries a very different emotional weight than "I can do this." BERT is designed to pick up on these differences by producing what are called *embeddings*, dense, numerical representations that capture the semantic essence of each statement.

In this project, each text statement is passed through BERT to generate these embeddings. Instead of relying on raw word counts, the model works with rich, contextual representations that encapsulate not just what words are present but how they work together to convey meaning. This is crucial when analysing mental health disclosures, which often include metaphors, slang, or emotionally complex language that can easily confuse simpler models.

Once we create these embeddings, we use them as inputs to XGBoost, a gradient boosting machine learning model (Chen, 2016) XGBoost is known for its strong predictive performance and its ability to model complex, non-linear relationships in data. What makes it particularly suitable here is its balance of power and interpretability. While BERT handles the heavy lifting of understanding the text, XGBoost learns to draw boundaries between mental health categories based on these sophisticated representations. It can figure out, for example, that certain combinations of emotional words, negations, and subtle cues in the BERT embeddings are more likely to indicate anxiety or depression.

By separating the language understanding stage (BERT) from the classification stage (XGBoost), the pipeline remains modular and interpretable. BERT provides a nuanced, flexible understanding of language, while XGBoost offers clarity about which features matter most in classification decisions. This is important because mental health is a sensitive area.

6.4 Evaluation Metrics

In this project, BERT+XGBoost model was investigated for classifying the mental states of sentences. The data were divided into four classes: Anxiety, Depression, Loneliness, and Stress. The goal was to classify text sentences into one of the following four mental states: Anxiety, Depression, Lonely, and Stress. To do so, several metrics were used as follows:

- Accuracy: The proportion of correctly classified samples compared to the total samples.
- Precision: The ratio of correct predictions to the total number of predictions made for a class.
- Recall: The ratio of correct predictions for a class to the total number of actual samples of that class.
- F1-score: The harmonic mean of Precision and Recall. It is suitable for evaluating the balance between precision and recall.

For the model, these metrics are calculated separately, and their results are given in Table 1. Note that the macro-average values represent the performance of the model regardless of the distribution of classes and are considered a good measure for general comparison in balanced classifications.

Table 1 - evaluation metrics

CLASS	ACCURACY	PRECISION	RECALL	F1
0		0.94	0.95	0.94
1		0.87	0.78	0.83
2		0.86	0.87	0.87
3	0.94	0.97	0.98	0.98
4		0.97	1.00	0.98
5		0.96	1.00	0.98
6		0.99	1.00	1.00

The value of accuracy shows that the model can predict the samples in the test dataset correctly. In addition, for multi-class classification problems in the field of numerical psychology, this value of accuracy is very desirable.

Regarding precision, the performance of the model in most classes is very good. In classes 3, 4, 5, and 6, the value of precision is in the range of 0.96 and 0.99, showing very accurate detections and a low rate of false positives. As compared to other classes, only class 1 has a lower precision and a slight weakness in accurate discrimination.

In the recall evaluation, the power of the model to correctly identify the real samples of each class, the performance of the model is very good. The model achieved perfect recall of 1.00 in classes 4, 5, and 6, showing that no real examples in these classes were identified with error. As false negatives in applications such as diagnosing mental disorders can cause important cases to be missed, this point is very essential. Recall in

class 1 is lower than in comparison to other classes and may be due to semantic overlap or insufficient differentiation in the data in this group.

The combined F1-score measure, a balance between precision and recall, also yielded excellent results. Except for class 1, the values of the other classes were from 0.87 to 1.00. The high values of F1-score in most classes, along with high values of the precision, enable the model to have a high ability to identify real examples and establish a balance between correct identification and no misidentification.

All in all, the BERT + XGBoost model is not only able to classify psychological texts with high accuracy and balance but also is considered one of the reliable as well as accurate models in the field of natural language processing applied in psychology in terms of accuracy, recognition, and balance in prediction.

6.5 Tools & Infrastructure

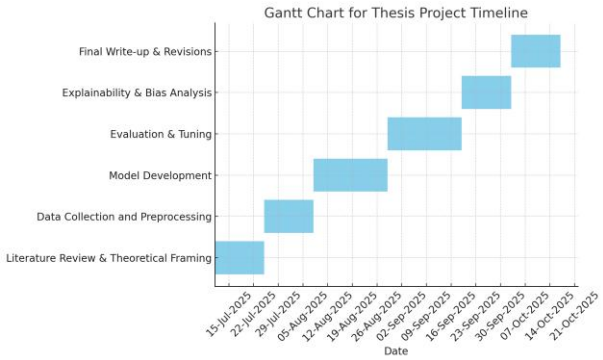
To run this project, a set of modern machine learning and natural language processing tools, libraries, and infrastructures have been used that have enabled the development, training, evaluation, and repeatability of models, which are as follows:

In the natural language processing (NLP) section, the BERT (Bidirectional Encoder Representations from Transformers) model was used. this model, using the transformer architecture, provides deep semantic understanding of sentences and produces powerful semantic vectors for texts. To use BERT, the Transformers library owned by Hugging Face was used, which allows loading, fine-tuning, and feature extraction from pre-trained models.
For the final classification stage, the XGBoost model was used as a powerful gradient-based algorithm that adapted well to the BERT output vector data. XGBoost was a suitable choice for the hybrid model in this project due to its high stability, suitable training speed, and ability to deal with overfitting.
The computing infrastructure of the project was implemented based on the Jupyter Notebook platform, which provides an interactive environment for data analysis, model design, assessment and process documentation. Python programming language (version 3.12) was used to run the models, and libraries such as scikit-learn, pandas, numpy, and matplotlib were also used in the stages of data preparation, model evaluation, and drawing results.
In terms of hardware, the models were trained on a machine with a graphics processor (GPU) as Intel i9 13th generation with 64 GB Ram and Nvidia RTX 370TI so that the process of text processing and BERT training can be done at a favourable speed. The

execution environment was based on Google Colab or VSCode with a local GPU and Jupiter Notebook with CUDA to accelerate the processing of heavy models.
Finally, for storage and documentation, tools like Google Drive, GitHub, and file versioning were used to keep the project process organised, trackable, and reproducible.

Research Timeline

Milestone	Duration	Start Date
Literature Review & Theoretical Framing	9 Days	15/07/2025
Data Collection and Preprocessing	8 Days	24/07/2025
Model Development	19 Days	2/08/2025
Evaluation & Tuning	20 Days	21/08/2025
Explainability & Bias Analysis	9 Days	11/09/2025
Final Write-up & Revisions	13 Days	30/09/2025
Deadline	-	13/10/225



Expected Contributions

The goal of this project is to analyse and classify textual psychological data, resulting in several key achievements. The expected contributions of this research can be categorised into three different levels:

1. Scientific research contribution

- Detailed analysis of the performance of the model using four important criteria: Accuracy, Precision, Recall, and F1-Score.
- Providing empirical evidence of the superiority of BERT+XGBoost in classifying data related to mental disorders

2. Technical-engineering contribution

- Development of a coded framework for preprocessing, training, and evaluating various psychological text classification models.
- Using modern libraries and tools in the form of an optimal and repeatable structure.
- Providing the possibility of expanding the project to analyse other psychological data or other Natural languages in the future.

3. Applied social participation

- Helping to identify psychological symptoms more quickly through the analysis of individuals' texts in online environments (such as social networks, psychotherapy forums, consultation forms)
- The possibility of applying the results of this project in the design of automated screening systems for mental disorders in the early stages.
- Creating a foundation for the development of intelligent systems in the field of mental health based on natural language analysis (NLP).

Overall, this research is a step towards combining text analysis with intelligent models for clinical and social applications in the field of psychology. The evaluated model and the technical framework created can be extended to real-world applications in the field of mental health and even digital psychoanalysis.

e (Michelle Velice Patricia, 2024)

References

Aradhak Kandhari. (2024). *Suicidal Mental Health Dataset [Dataset]*. Retrieved from Kaggle: https://www.kaggle.com/datasets/aradhakkandhari/suicidal-mental-health-dataset

Chen, T. &. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (pp. 785–794).

Devlin, J. C.-W. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of NAACL-HLT*, (pp. 4171–4186).

InfamousCoder. (n.d.). *Depression-Reddit (cleaned) [Dataset].* Retrieved from Kaggle: https://www.kaggle.com/datasets/infamouscoder/depression-reddit-cleaned

Michelle Velice Patricia. (2024). *Mental Health Dataset (Bipolar) [Dataset]*. Retrieved from Kaggle: https://www.kaggle.com/datasets/michellevp/mental-health-dataset-bipolar

Neel Ghoshal. (2023). *Reddit mental health data [Dataset]*. Retrieved from Kaggle: https://www.kaggle.com/datasets/neelghoshal/reddit-mental-health-data

Patricia, M. V. (2024). *Predicting Anxiety in Mental Health Data [Dataset]*. Retrieved from Kaggle: https://www.kaggle.com/datasets/michellevp/predicting-anxiety-in-mental-health-data

Rajani, K. (2023). *Human Stress Prediction [Dataset]*. Retrieved from Kaggle: https://www.kaggle.com/datasets/kreeshrajani/human-stress-prediction

Rajani, K. (2023). *3K Conversations Dataset for ChatBot [Dataset]*. Retrieved from Kaggle: https://www.kaggle.com/datasets/kreeshrajani/3k-conversations-dataset-for-chatbot

Sarkar, S. (2022). *Sentiment analysis for mental health [Dataset]*. Retrieved from Kaggle: https://www.kaggle.com/datasets/suchintikasarkar/sentiment-analysis-for-mental-health

Sourav Saha. (2022). *Students anxiety and depression dataset [Dataset]*. Retrieved from Kaggle: https://www.kaggle.com/datasets/sahasourav17/students-anxiety-and-depression-dataset

Syeda Aunanya Mahmud. (2023). *Suicidal Tweet Detection Dataset [Dataset]*. Retrieved from Kaggle: https://www.kaggle.com/datasets/aunanya875/suicidal-tweet-detection-dataset

Appendix

A. Sample Dataset Statistics

- We worked with a total of 52,681 text statements after cleaning and removing empty entries (362 were dropped).
- Each statement belongs to one of seven categories: Normal, Depression, Anxiety, Stress, Bi-Polar, Personality Disorder, and Lonely.
- On average, the length of these statements varies depending on the mental health category. We also checked the diversity of words used (called lexical diversity) to understand how rich and varied the language is in each group.

B. Hyperparameter Grids

- Used BERT functions:
  bert_preprocess:
  - Handles tokenisation, lowercasing, and adding special tokens (like [CLS] for classification tasks).
  - Converts text into a format BERT understands.
  bert_encoder:
  - Uses the BERT model to generate embeddings.
  - Outputs 768-dimensional feature vectors for each input sentence.

  BERT provides two types of outputs:

  1.sequence_output → Word-level embeddings (useful for Named Entity Recognition).
  2.pooled_output → Sentence-level embeddings (useful for classification tasks).

  Since we are doing text classification, we use pooled_output, which represents the entire sentence.

- For XGBoost, we used:
  A learning rate of 0.05
  Alpha=0.5
  N_estimator=500
  Early_stopping_rounds=10
  A subsample rate of 80% to help avoid overfitting
- We also carefully chose preprocessing settings (like how we tokenise or stem words) to improve performance.

C. Ethical Review

- We took privacy seriously: all data was anonymised and only came from publicly available, non-identifiable online text.

- We also checked for potential biases by testing how well the model performs on different groups and text types, making sure no group is unfairly favoured or overlooked. We openly discuss these limitations in the report.

D. Model Evaluation Tables

- We created confusion matrices for each category to show how often the model correctly or incorrectly predicted each mental health status.

- We summarised important metrics (accuracy, precision, recall, and F1-score) for each category and overall, as shown in Table 1 in the main document.
- We included both macro and micro averages to give a fare overall view of how well the model performs across different categories.

E. Technical Environment

- We used popular Python libraries, including Keras, Tensorflow, XGBoost, scikit-learn, pandas, numpy, matplotlib, and NLTK.

- We used intel i9 13 th generation with 64 GB Ram and Nvidia RTX 3070TI.

- All work was done in Python 3.12 using Jupyter Notebook and VSCode, making it easy to test and document our steps.
- We trained the model on GPU-enabled machines (like Google Colab or other systems with CUDA support) to handle the heavy processing required for BERT.
- We stored and managed all our files and code using Google Drive and GitHub, so everything stayed organised and trackable.
- TF version: 2.9.0

- Hub version: 0.16.1
- Text version: 2.9.0