Can Words Reveal Mental Health? A Machine Learning Approach

Introduction

Mental health issues such as depression or anxiety become increasingly prevalent (World Health Organization (WHO), 2022), yet many of them go undetected due to reasons such as stigma, lack of awareness or limited access to professional help (Merced, 2023). As digital communication on social media platforms become more and more embedded in daily life (Chaffey, 2024), there is a growing opportunity to identify early signs of mental health issues through language, potentially leading to early and successful intervention.

Based on this motivation, this project aims to explore whether machine learning models can accurately predict an individual's mental health status based on the content of their written statements. By analyzing text data utilizing Natural Language Processing (NLP) and a range of machine learning (ML) techniques, I aim to detect mental health conditions such as depression, anxiety, or bi-polar disorder and others, given a text input. The central research question guiding this project is: *Can text-based features from an individual's written statements be used to accurately predict their mental health status using machine learning techniques?*

Dataset

The data used in this project was created by Suchintika Sarkar and obtained from Kaggle. It compiles text data from publicly available datasets and sources, including social media posts, Reddit posts, Twitter posts etc. The dataset consists of 54,043 statements, each labeled with one of the following mental health statuses: depression, anxiety, normal, stress, suicidal, bipolar, personality disorder.

Scope of the Project

The project follows a structured pipeline with three following key phases: data preprocessing, exploratory data analysis (EDA) and model building and evaluation. The Data Preprocessing stage consists of text cleaning (removing punctuation, lowercasing, etc.), handling missing values and duplicates, applying tokenization and lemmatization using NLTK, and transforming text into numerical representations using TF-IDF vectorization. Next, exploratory data analysis is conducted to gain insights into the distribution of mental health categories, word frequencies across mental health categories, as well as relationships

between features and mental health statuses. Lastly, multiple traditional classification algorithms are trained and evaluated, including: logistic regression, naive bayes, support vector machine (SVM), and XGBoost. Additionally, a feedforward neural network is applied using PyTorch in order to evaluate whether deeper architectures improve classification performance. Model performance is assessed using accuracy, F1 scores, and confusion matrices to understand how well each model captures the nuances of mental health categories.

Conclusion

The results show that machine learning models can achieve a reasonably high level of accuracy in predicting mental health status based on text data. Among all models tested, XGBoost emerged as the top performer, delivering the highest accuracy and F1 scores across most classes, especially in handling both common and underrepresented mental states. Surprisingly, simple ML models like Logistic Regression and SVM outperformed deeper neural networks, which may suggest that for sparse, high-dimensional TF-IDF vectors, traditional algorithms remain highly effective and efficient. The deep learning models, while promising in theory, struggled to outperform classical approaches. This underperformance may stem from a few reasons such as TF-IDF inability to capture semantic and grammatical context or minimal tuning of hyperparameters.

Limitations of the project

One of the limitations of the project is that it relied on TF-IDF vectorization. While TF-IDF offers a nuanced representation combining term frequency with inverse document frequency, it ignores word order, syntax or context (Jain, 2024). Thus, important linguistic nuances or emotional tone may be lost, limiting the model's ability to understand and interpret. This could also be a likely explanation for why neural networks did not outperform basic models.

Secondly, the implemented neural networks models were rather simple and had limited hyperparameter tuning. The classification could benefit from more advanced architectures such as transformers which could be suited better for modeling the language, potentially improving performance.

Lastly, while RandomOverSampler was used to address class imbalance, it is not without weakness. To make the classes balanced, it duplicated minority classes examples

synthetically, which could lead to potential overfitting and overoptimistic performance, particularly for rare categories such as personality disorder.

Future Work

In future work, the limitations from the section above could be addressed. Therefore, another method for text analysis could be experimented e.g., embeddings like BERT or RoBERTA, and other resampling techniques could be tested. Moreover, different neural networks architectures could be taken into account.

Additionally, I consider simplifying the multi-class classification problem into a binary classification problem, which would involve distinguishing between text statements of individuals showing signs of any mental health disorder versus those without it. This idea emerged from observing significant overlap in the language across some categories e.g., depression and suicidal. Moreover, many symptoms for mental health disorders overlap, which also could be reflected in text statements. By aggregating all disorder categories into one class, the model might learn to better capture general patterns of psychological disorders. This could potentially lead to better overall performance and could serve as a useful early-warning tool for identification of individuals who could benefit from further screening or professional support.

References

Chaffey, D. (2024, May). *Global social media statistics research summary 2024 [May 2024]*. Smart Insights. https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/?utm_source=chatgpt.com

Jain, A. (2024, February 4). *TF-IDF in NLP (Term Frequency Inverse Document Frequency)*. Medium. https://medium.com/@abhishekjainindore24/tf-idf-in-nlp-term-frequency-inverse-document-frequency-e05b65932f1d

Merced, A. D. L. (2023, August 21). *Facts & Figures About Undiagnosed Mental Disorders*. Remedy Psychiatry, Inc. https://remedypsychiatry.com/facts-figures-about-undiagnosed-mental-disorders/

World Health Organization (WHO). (2022, June 8). *Mental Disorders*. World Health Organization; World Health Organization. https://www.who.int/news-room/fact-sheets/detail/mental-disorders