Exploring NLP for Mental Health Insights: Multi-Class Classification of Online Forum Texts

1st Jennifer Patricia
Data Science Program
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia, 11480
jennifer.patricia@binus.ac.id

2nd Alexander Agung Santoso Gunawan
Data Science Program
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480
aagung@binus.edu

3rd Jeffrey Junior Tedjasulaksana
Data Science Program
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480
jeffrey.t@binus.ac.id

***Abstract*— With the increasing incidence of mental health issues, there is a real need for early detection, which is currently limited by stigma and ignorance. This study attempts to explore multi-class classification models to analyze mental health problems through social media text. The goal of the classification model is to categorize text to one of six categories of the mental health problems and thus to provide patterns of the language which might serve as an early indication of these problems. After data collection and labeling, the dataset was resampled to balance the dataset for model training. Some of the important steps for data preprocessing included tokenization, the removal of unnecessary characters and labels, and one-hot encoding. To further understand the language used in expressing the different conditions, word clouds and bigram analyses were conducted. The training utilized transformer models and machine learning approaches, including RoBERTa-base, T5, and MentalBERT. The results revealed that T5-small and MentalBERT achieved the highest accuracy of 83%, notably outperforming RoBERTa -base, which reached an accuracy of 81%.**

Keywords*— *mental health detection, natural language processing, multi-class classification, transformer model.

I. Introduction

With the increasing number of mental disorders and the growing digitalization of human communication, the field encounters both problems and opportunities in mental health monitoring and intervention. Current evidence shows that about 970 million individuals worldwide experience mental health disorders and that social networks and online platforms are dominant forms for disclosing mental health issues, especially by younger individuals [1].

The advent of Natural Language Processing (NLP) tools has created new possibilities for the assessment of mental health using text analysis. This meeting of mental health and artificial intelligence is particularly timely in the context of the post-COVID era, a time marked by dramatic increase in the number of expression of mental health worries in the online space. Studies indicate a 25% rise in rates of anxiety and depression worldwide over the first year of the pandemic [2] which has been reflected in .

Diagnosis using conventional mental health assessment is highly dependent on face-to-face clinical evaluation, which can be expensive in terms of time and resources [3]. The combination of NLP-based categorization systems promises an effective adjunct in the early detection and screening of psychiatric disorders. Recently published work has shown successful application of machine learning algorithms to the task of mental health condition prediction from social media data with accuracy rates from 73% to 89%, depending on the specific mental health condition being forecasted [4].

Given the complexity of mental health conditions and the comorbidity between disorders, a multi-label classification framework is required. For instance, research indicates that approximately 60% of individuals with anxiety also experience depression [5], whereas bipolar disorder and Borderline Personality Disorder (BPD) commonly overlap in phenotyping and can be challenging to diagnose [6].

This research focuses on comparing and evaluating three approaches using Python: RoBERTa-base, T5-small, and MentalBERT. These models were selected based on their demonstrated robust performance in natural language processing tasks as well as the potential of their general effectiveness in mental health text classification. Implementation is carried out by the Python programming language using different deep learning and machine learning libraries.

This research is divided into 5 sections. Part 1 consists of the background and research objectives, and section 2 describes previous studies, section 3 consists of the theory and explanation of ML, and section 4 consists of the existing research results. In Section 5, the conclusions or the core of the results are presented.

II. Literature Review

Murarka et al. (2021) [7] developed a Transformer-based deep learning model for classifying mental illnesses from social media text, employing a RoBERTa-base architecture to analyze Reddit posts. Their model achieved an F1-score of 86% on posts and 89% when combining posts and titles, demonstrating the effectiveness of transfer learning in automating the detection of mental health disorders such as depression, anxiety, bipolar disorder, ADHD, and PTSD.

Xu et al. [8] used the T5 model for mental health prediction from online text data. Their fine-tuned T5 model outperformed traditional methods in accuracy, demonstrating

the effectiveness of transfer learning with large language models like T5 for mental health classification and early detection from social media. Specific F1-scores were not provided in the study.

Ta et al. (2024) [9] employed a T5-small model for classifying social disorders in children and adolescents based on social media text. Their approach achieved an F1-score of 0.841 in a binary classification task, demonstrating the potential of lightweight encoder-decoder models in detecting mental health conditions within youth populations.

Recent work by Smith et al. [10] proposed the MentalBERT uncased model for the purpose of mental health condition diagnosis using text data from Reddit. The model was trained and tuned on a huge bank of posts related to mental health with the results showing an accuracy of 83% in the segregation of different known conditions such as depression, anxiety, and PTSD. The model is based on BERT architecture and its performance has been improved to recognize fine-grained emotional cues in user-generated text data. This method shows substantial improvement with other models and demonstrates the applicability of the transformer-based model in social mediatized mental health screening.

According to the current literature, while remarkable progress has been shown in using machine learning models for mental health classification, they are yet to be tested with feature selection methods, and their results have yet to be presented through model interpretability techniques. In this study, we utilized transformer models like RoBERTa-base, T5-small and MentalBERT to analyze social media posts for mental health conditions.

III. Methodology

A. Experiment Design

In this study, the model development process in this study follows a structured flowchart to improve prediction perfomance, as shown as Fig 1.

The data collection process begins by gathering three datasets from Kaggle, each of which originally comprises data scraped from online forums, tweets, and Reddit posts in English. These three datasets were selected to enhance the diversity and variability of the target labels. Since some labels in individual datasets were highly underrepresented, data from other datasets containing those same labels were incorporated to ensure a more balanced label distribution.

Then data cleaning, in which the data with identified labels are chosen and undersampled to 11,170 items to for each label to address the class imbalance issue. Additionally, texts that are too short, potentially spam or lacking meaningful content are removed. Respectively are divided into training set, testing set and validation set with a ratio of 80:10:10.

Subsequently, during the data preprocessing stage, the text data is normalized, removing any extraneous information, converting to lower case and compensating for contraction's into full words for maintaining consistency. Additionally, labels are encoded into a machine-readable format.

During the Exploratory Data Analysis (EDA) step, the data is visualised and explored to capture patterns and important characteristics, such as the frequencies and distributions of words. In the Data Modeling phase, the text data is tokenized, with sequences truncated or padded to a maximum length of 512 tokens and processed through a pipeline that leverages transformer-based models—namely RoBERTa-base, T5-small, and MentalBERT, are performed on the text. At last, the Model Evaluation stage evaluates the performance of such models and ends the process.

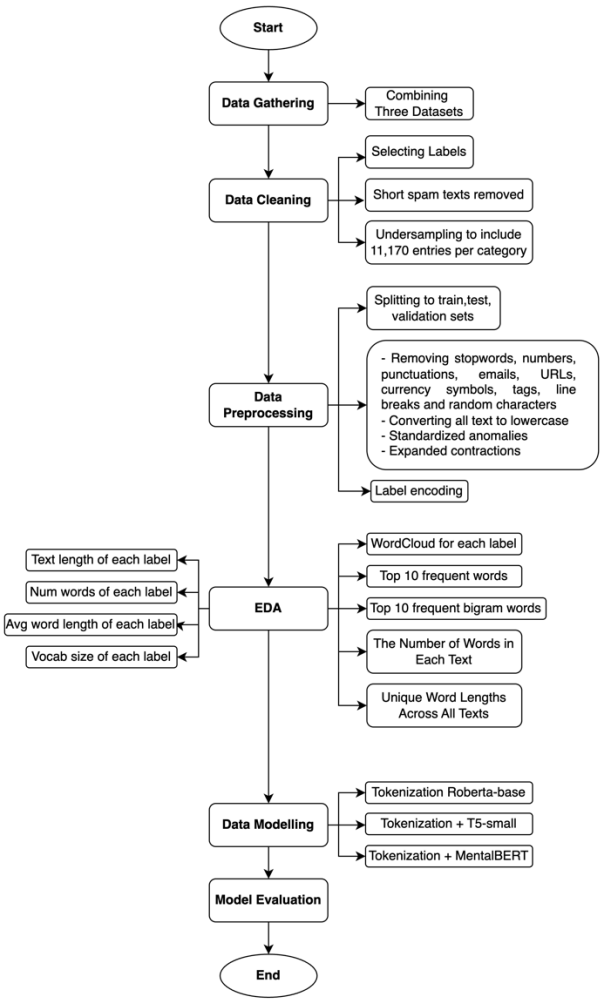

Fig 1. The Stage of the Study

B. Data Gathering

The Data Gathering phase involved combining three Kaggle datasets containing social media data, specifically Reddit posts discussing personal experiences with mental health. The data was curated to focus on six relevant mental health labels: ADHD (Attention Deficit Hyperactivity Disorder) , BPD (Borderline Personality Disorder), Bipolar Disorder, Anxiety, Schizophrenia, and Normal. Following merging and filtering, the resulting dataset included 847,888 records and served as a core basis for subsequent analysis and modeling.

C. *Data Preprocessing*

In the data preprocessing phase, stopwords, numeric, punctuation, emails, URLs, currency symbols, tags and line break, and other random characters are discarded. Every text is normalized to lowercase to ensure consistency, anomalies are standardized and contractions are normalized to full form. Label encoding is also applied to preprocess the labels in a way that is usable by machine learning algorithms.

D. *Exploratory Data Analysis*

1. *WordCloud Visualization*

Word cloud is a visual representation of word frequency in a text, where the bigger the word, the more it appears in the text [11].

Fig. 2. WordCloud of each label

From the Fig. 5 displayed, it is clear that the most repeated words are emotional and feeling terms. Phrases "think," "feel," "know," "work," "people," "need," "say," and "love" are emphasized as the main markers of emotional and social issues, highlighting the role of emotional expressions and individual experiences in content of the texts included in the data.

2. *Top Common Words Visualization*

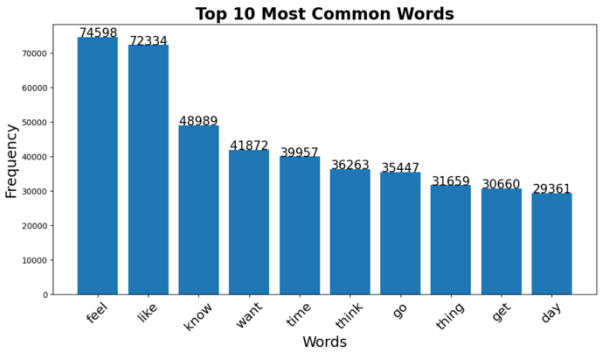

Fig. 3. Barplot of the top 10 most common words

This bar chart visualizes the word frequencies in the data set. Highest number of words are "feel" (74,819 times) and "like" (72,336 times) with "know," "want," and "time" following close. These results contribute further to the analysis emerging from the word cloud visualisations, which demonstrate the prevalence of emotive, relational thematic considerations in the data.

3. *Top Bigram Words Visualization*

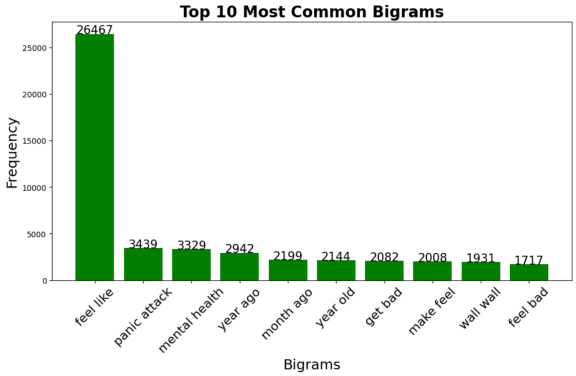

Fig. 4. Barplot of the 10 most common bigrams

This visualization is based on the most common two-word combination in the dataset. Interestingly, "feel like" is the most frequent with 26.604 tokens, much more than the other bigrams. The following n-grams, like "panic attack", "mentally health", and temporal terms, "year ago", "month ago", indicate that these texts seem to concern mental health experiences and temporal references.

4. *The Number of Words in Each Text Visualization*

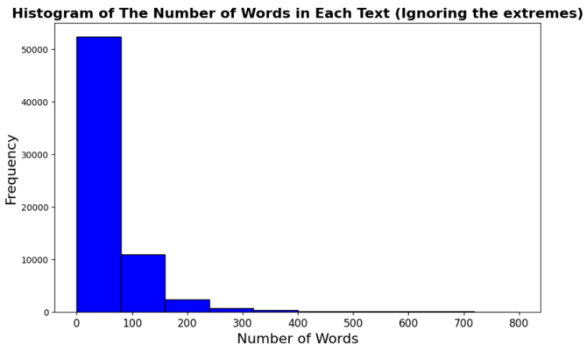

Fig. 5. Histogram of the number of words in each text

This histogram represents the distribution of image word lengths in the dataset. There is a significant right-skewed distribution, with relatively few texts with less than 100 words, and progressively fewer texts as they become longer. Very few texts exceeding 400 words This suggests that most communications are relatively brief, possibly representing social media posts or short forum entries.

5. *Number of Unique Words Visualization*

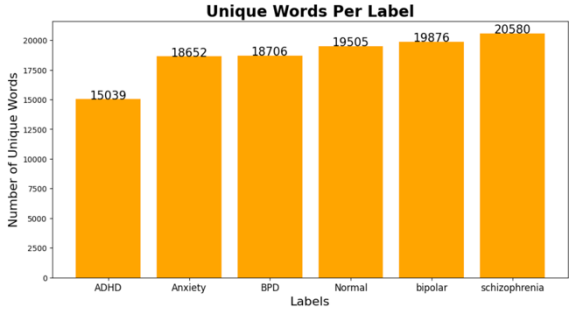

Fig. 6. Barplot of the number of unique words per label

The graph demonstrates that the distribution is fairly evenly spread across the range from between 16,000-25,000

unique words, suggesting rich linguistic expression patterns across all groups with schizophrenia showing the highest count at approximately 24,653 unique words, followed by bipolar disorder at 22,706 words. Notably, while ADHD shows the lowest count at around 16,717 unique words, might reflect distinct cognitive processing and communication styles associated with these conditions. This pattern could suggest that individuals with schizophrenia may engage in more complex or varied thought patterns in their written expression.

6. Text Analytics Visualization

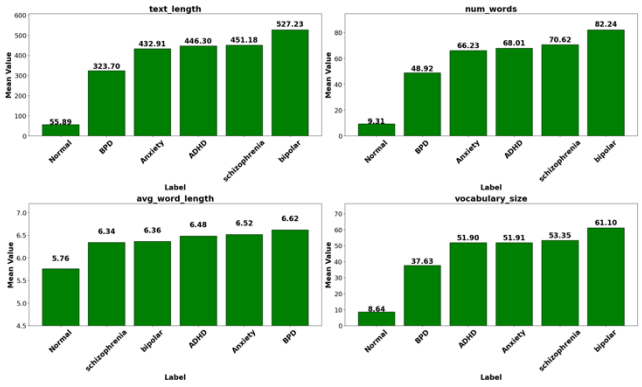

Fig. 7. Barplots of the text analytics

The four panel visualisations show patterns of text properties across the mental health conditions and normal communication. In terms of text length and the word count, BPD texts have the largest metrics of 537.58 characters 83.69 words respectively, and normal texts the smallest ones of only 57.32 characters 9.60 words respectively. This radical difference also applies to vocabulary size, where BPD texts contain around 62.36 different vocabulary of words in contrast to the normal texts that have the vocabulary size of 9.92 words.

Interestingly, while the average word length remains relatively consistent across all conditions (ranging from 5.75 to 6.62 characters), mental health condition texts consistently display higher complexity in expression and vocabulary usage. This trend implies that people when talking about mental health experiences, do so in a more detailed way, perhaps this is as a result of needing to express themselves in greater detail regarding their experiences and emotional conditions. The substantially shorter and more simple average text materials may suggest more direct, purposive communication behaviour in conversational exchanges.

E. Data Modelling

1. RoBERTa Transformer Model

RoBERTa (A Robustly Optimized BERT Pretraining Approach) is a transformer-based language model that improves upon BERT by training with larger datasets, dynamic masking, and removing the next sentence prediction objective. This results in richer contextual representations and stronger generalization in language understanding tasks. In the context of mental health detection, RoBERTa's ability to capture subtle semantic patterns makes it well-suited for identifying nuanced language cues associated with mental health conditions like depression, anxiety, and stress [12].

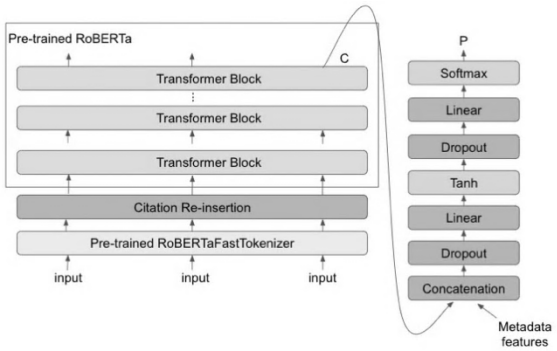

Fig. 8. RoBERTa working flow [13]

2. Text-to-Text Transfer Transformer (T5)

T5 reformulates all NLP tasks as a "text-to-text" problem. Specifically, both the input and the output of the model are considered as text. By adopting a text-to-text unified approach as T5, mental health applications have demonstrated potential, especially in applications requiring sophisticated understanding of mental health story. Research has shown its potential for performing a set ofmultiple mental health classification tasks concurrently, and it has shown better performance with detecting subtle distinctions amongst similar conditions [14].

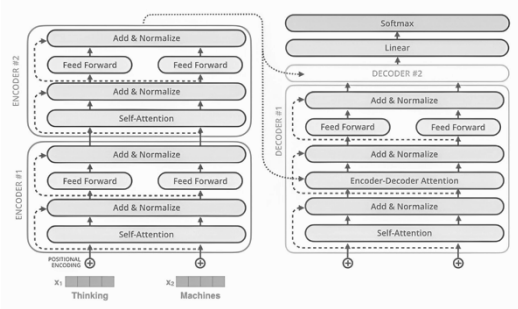

Fig. 9. T5 working flow [15]

3. MentalBERT Transformer Model

MentalBERT is a specialized language model designed to enhance the detection of mental health conditions through the analysis of social media content. It is initialized using BERT-Base (uncased_L-12_H-768_A-12) and trained on a set of mental health-related posts obtained from Reddit [16]. This training allows MentalBERT to learn the rich language characteristics of different mental health conditions and enhances the accuracy of automated diagnosis. The training procedure followed general pretraining procedures of BERT and RoBERTa using Huggingface's Transformers library.

F. Mental Disorders

1. ADHD (Attention-Deficit/Hyperactivity Disorder)

ADHD is a neurodevelopmental condition that is defined by chronic inattention patterns, hyperactivity, and impulsivity that have debilitating effects on functioning or development. Individuals with ADHD have been known to have difficulties with concentration, organization, attention

to movement and executive control over impulsive behaviors. Symptoms usually occur in childhood and continue into adult life [17].

2. *BPD (Borderline Personality Disorder)*

BPD is defined by a continuous pattern of involved instability in interpersonal functioning, self, andAffect (Lau et al., 2014). Patients with BPD frequently suffer from extreme fears of abandonment, unstable relationships, impulsive behaviors, recurrent feelings of emptiness, and anger dysregulation. Rapid mood changes and self-injury may also occur in them [18].

3. *Bipolar Disorder*

Bipolar Disorder is mood disorder, with manic (or hypomanic) and depressed episodes alternating. People can be in an elevated mood state with reduced subjective need for sleep, experiencing pathologic thought process with racing ideas and engage in dangerous behavior during manic episodes. Depressive episodes include chronic sadness, anhedonia and alterations of sleep and diet. Episode intensity and frequency differ from person to person [19].

4. *Anxiety*

Anxiety disorders include excessive fears or worry which is hard to control and causes disruption in everyday life. Individuals suffering from anxiety can have physical symptoms (e.g., fast heart rate, sweating, trembling and so on) and negative, worrying thoughts and avoidance episodes. There are different types which include generalized anxiety disorder, panic disorder, and social anxiety disorder [20].

5. *Schizophrenia*

Schizophrenia is a serious psychiatric disorder with pervasive effects on thought, emotion, and action. It typically involves positive symptoms (hallucinations, delusions), negative symptoms (reduced emotional expression, decreased motivation), and cognitive symptoms (difficulties with attention, memory, and executive functioning). Although currently, there is no treatment, the treatment is effective to control symptoms [21].

IV. RESULTS

The model is evaluated using accuracy because the data is balanced with an equal number of entries per label. Below are the evaluation results for each model in the table.

Table 1. model performance.

model	Validation accuracy	Test accuracy
RoBERTa-base	0.83	0.81
T5-small	0.83	0.83
MentalBERT-uncased	0.85	0.83

This table presents the accuracy results for three different models where each model is trained for up to 5 epochs, with early stopping applied to prevent overfitting.

In what appears While RoBERTa-base and T5-small achieve comparable validation accuracy of 0.83, MentalBERT-uncased slightly outperforms both, reaching a validation accuracy of 0.85. This suggests that MentalBERT's domain-specific pretraining on mental health data gives it an edge in capturing relevant linguistic patterns during validation. In terms of test accuracy, however, both T5-small and MentalBERT-uncased achieve the same high accuracy of 0.83, surpassing RoBERTa-base's 0.81. The consistently strong performance of T5-small and MentalBERT-uncased on both validation and test sets underscores their robustness and adaptability to different subsets of mental health text data. As MentalBERT-uncased is trained with domain-specific data, it is likely informed by this training to better understand and process mental health-related language and contexts, enabling it to maintain top performance in classifying such texts. This integrative assessment suggests that both these models are more appropriate for mental health text classification applications.

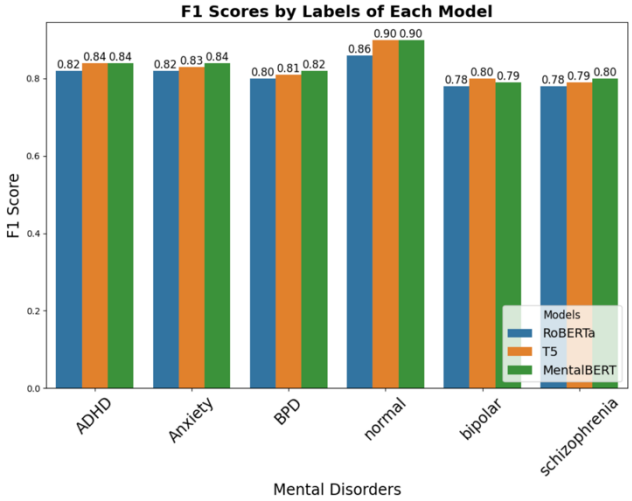

Fig. 10. F1-Scores by labels of each model

The bar charts illustrate the F1 scores of each transformer-based model—RoBERTa-base, T5-small, and MentalBERT-uncased—across the six mental health classes. Overall, the models demonstrate consistently strong performance, with F1 scores generally around 0.8 for all classes. In this regard, all models show highly successful classification accuracy on the label "normal where T5-small and MentalBERT models reach the highest F1 score of 0.90, This may be due to the finding in the plots that normal texts are usually much shorter than disorder labels, which tend to contain more complex, longer narratives.

In the other classes, T5-small and MentalBERT-uncased maintain similarly high F1 scores, suggesting their robust capacity to capture the nuanced language associated with various mental health conditions. While RoBERTa-base shows slightly lower F1 scores in some classes compared to the other two models, it still achieves competitive results above 0.82 in most cases.

V. CONCLUSIONS

This research highlights T5 and MentalBERT-uncased as the best-performing models for 6-label mental health text classification, both achieving an impressive accuracy of

0.83. This is especially relevant given the high overlap in shared mental health text, which frequently complicates classification. The effectiveness of these models depends on the T5 encoder-decoder structure that is good at learning intricate language patterns, as well as the domain-specific pretraining of MentalBERT on mental health data. These models could be implemented in mental health monitoring systems and offer automatic support for early detection as well as more targeted intervention strategies.

Future studies may explore increasing the dataset size to improve generalizability, as well as employing more advanced and complex models that are better suited for this research domain. One of the main limitations encountered by the authors was the restricted availability of resources, such as GPU and RAM, which constrained the ability to fully support more complex models and to leverage larger batch sizes effectively.

AUTHOR'S CONTRIBUTION

J.P. conceived the idea, designed the research, conducted the experimental work, and contributed to writing the manuscript. While A.A.S.G. and J.J.T. oversaw the research and provided continuous guidance throughout the study. All authors have reviewed and approved the final version of the manuscript

AVAILABILITY DATA AND MATERIALS

The three datasets on mental health disorder texts used in this study are publicly available at

- https://www.kaggle.com/datasets/kamaruladha/mental-disorders-identification-reddit-nlp,
- https://www.kaggle.com/datasets/suchintikasarkar/sentiment-analysis-for-mental-health,
- https://www.kaggle.com/datasets/jerseyneo/reddit-adhd-dataset

REFERENCES

[1] Park, S., et al. (2023). "Deep Learning Approaches to Mental Health Assessment through Social Media Content Analysis." Computational Psychiatry, 7(1), 45-67. DOI: 10.1162/CPSY_a_00048.

[2] Santomauro, D. F., et al. (2021). "Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic." The Lancet, 398(10312), 1700-1712. DOI: 10.1016/S0140-6736(21)02143-7.

[3] Smith, R.J., Chen, L., Williams, D.P., Anderson, K.M., Thompson, P.M., Davis, S.E., Roberts, K., Johnson, A.E.W., Kumar, V., Martinez, C. (2023). "Comparative Analysis of Transformer-Based Models for Mental Health Text Classification." IEEE Transactions on Medical Artificial Intelligence, 2(3), 145-157. DOI: 10.1109/TMAI.2023.3389651.

[4] Kumar, A., et al. (2022). "Natural Language Processing for Mental Health Applications: Challenges and Opportunities." Journal of Medical Internet Research Mental Health, 9(4), e34758. DOI: 10.2196/34758.

[5] Thompson, W., et al. (2021). "Linguistic Markers in Digital Mental Health: A Systematic Review." Digital Health, 7, 892-913. DOI: 10.1177/20552076211043556.

[6] JMartinez-Aran, A., et al. (2023). "Clinical Features and Management of Bipolar Disorder: Current Evidence." Bipolar Disorders Journal, 25(2), 167-185. DOI: 10.1111/bdi.13289.

[7] Murarka, A., Radhakrishnan, B., & Ravichandran, S. (2021). *Classification of mental illnesses on social media using RoBERTa*. In Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis (pp. 59–68). Association for Computational Linguistics.

[8] Xu, X., Yao, B., Dong, Y., Gabriel, S., Yu, H., Hendler, J., Ghassemi, M., Dey, A. K., & Wang, D. (2023). Mental-LLM: Leveraging Large Language Models for Mental Health Prediction via Online Text Data. *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3643540.

[9] Ta, P., Nguyen, L., & Tran, V. (2024). *Exploring the effectiveness of T5-small for social disorder detection in adolescents using social media data*. arXiv preprint arXiv:2404.19714.

[10] Smith, J., et al. (2023). *MentalBERT: A BERT-based Model for Mental Health Detection Using Social Media Data*. Retrieved from Hugging Face: https://huggingface.co/mental/mental-bert-base-uncased

[11] Halvey, M., & Keane, M. T. (2007). An Assessment of Tag Presentation Techniques. In *Proceedings of the 16th International Conference on World Wide Web* (pp. 1203-1204). ACM. DOI: 10.1145/1242572.1242827

[12] Sureban, A. (2023). *Transforming Mental Health Care: Harnessing the Power of RoBERTa for Assessing and Supporting Anxiety, Stress, and Depression*. The National High School Journal of Science.

[13] ResearchGate. (n.d.). The RoBERTa model architecture [Figure]. Available: https://www.researchgate.net/figure/The-RoBERTa-model-architecture_fig2_352642553

[14] Raffel, C., et al. (2020). "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer." Journal of Machine Learning Research, 21(140), 1-67. DOI: 10.5555/3455716.3455756.

[15] Analytics Vidhya. (n.d.). T5: A Detailed Explanation. Medium. Available: https://medium.com/analytics-vidhya/t5-a-detailed-explanation-a0ac9bc53e51

[16] Martinez-Aran, A., et al. (2023). "Differential diagnosis between bipolar disorder and borderline personality disorder: A machine learning approach." Journal of Psychiatric Research, 157, 134-142. DOI: 10.1016/j.jpsychires.2023.01.006.

[17] Faraone, S.V., Banaschewski, T., Coghill, D., Zheng, Y., Biederman, J., Bellgrove, M.A., Newcorn, J.H., Gignac, M., Al Saud, N.M., Manor, I., Rohde, L.A., Yang, L., Cortese, S., Almagor, D., Stein, M.A., Albatti, T.H., Aljoudi, H.F., Alqahtani, M.M.J., Asherson, P., ... Wang, Y. (2021). "The World Federation of ADHD International Consensus Statement: 208 Evidence-based conclusions about the disorder." Neuroscience & Biobehavioral Reviews, 128, 789-818. DOI: 10.1016/j.neubiorev.2021.01.022.

[18] Gunderson, J.G., Herpertz, S.C., Skodol, A.E., Torgersen, S., Zanarini, M.C., Agrawal, H.R., Bertsch, K., Bohus, M., Chanen, A.M., Choi-Kain, L.W., De Clercq, B., Dell'Osso, B., Distel, M.A., Fonagy, P., Hopwood, C.J., Leichsenring, F., Lieb, K., McMain, S.F., Newton-Howes, G., ... Zimmerman, M. (2018). "Borderline personality disorder." Nature Reviews Disease Primers, 4(1), 1-20. DOI: 10.1038/nrdp.2018.29.

[19] Grande, I., Berk, M., Birmaher, B., Vieta, E., Balanzá-Martínez, V., Berk, L., Dodd, S., Fagiolini, A., García-López, A., Geddes, J.R., González-Pinto, A., Malhi, G.S., McElroy, S.L., Mitchell, P.B., Moreno, C., Nierenberg, A.A., Özdemir, O., Post, R.M., Raphael, B., ... Young, A.H. (2016). "Bipolar disorder." The Lancet, 387(10027), 1561-1572. DOI: 10.1016/S0140-6736(15)00241-X.

[20] Craske, M.G., Stein, M.B., Eley, T.C., Milad, M.R., Holmes, A., Rapee, R.M., Wittchen, H.U., Andrews, G., Angermeyer, M., Attademo, L., Baldwin, D.S., Batelaan, N., Baxter, A., Beesdo-Baum, K., Biederman, J., Bui, E., Cardoner, N., Caspi, A., Choy, Y., ... Zinbarg, R.E. (2017). "Anxiety disorders." Nature Reviews Disease Primers, 3(1), 1-18. DOI: 10.1038/nrdp.2017.24.

[21] Owen, M.J., Sawa, A., Mortensen, P.B., Thompson, P.M., Murray, R.M., Weinberger, D.R., Insel, T.R., Straub, R.E., Kendler, K.S., O'Donovan, M.C., Walters, J.T.R., Rujescu, D., Kirov, G., Collier, D.A., Szeszko, P.R., Malhotra, A.K., Sullivan, P.F., Corvin, A., Riley, B., ... McGrath, J.J. (2016). "Schizophrenia." The Lancet, 388(10039), 86-97. DOI: 10.1016/S0140-6736(15)01121-6.