Mental Health Trigger Detection for Timely Intervention

Poulami Ghosh
Graduate Student
Department of Data Science
Rochester Institute of Technology
pg2360@rit.edu

Abstract

This study investigates the use of sentiment analysis to identify mental health triggers in user-generated content. The basic purpose is to categorize text into mental health-related categories and identify probable stressors connected with each. Using advanced natural language processing (NLP) approaches such as fine-tuned transformer models and zero-shot classification, this research achieves considerable accuracy in mental health predictions. It also discusses crucial issues including imbalanced datasets and the complexities of nuanced sentiment analysis. The data are provided as detailed visualizations, providing actionable insights to inform demographic mental health studies and tailored interventions.

1 Introduction

Mental health concerns are one of today's most important public health challenges, affecting individuals and communities around the world. With the growth of social media and online forums, user-generated information has become an invaluable resource for interpreting public emotion and spotting possible mental health triggers. These internet platforms provide information on people's thoughts, feelings, and pressures, making them critical for early detection and intervention. Traditional approaches to mental health analysis usually rely on surveys and clinical observations(Suominen, 2014), which are both time-consuming and limited in scope. By classifying text into predefined mental health categories, such as Anxiety, Bipolar, Depression, Normal, Personality disorder and Stress, and identifying associated stressors, this study seeks to enhance the understanding of mental health dynamics in real-time. The research emphasizes the application of advanced NLP methods, such as fine-tuned transformer models and zero-shot classification, to obtain high accuracy in mental health predictions. It also explores the intricacies of nuanced sentiment analysis and the challenges given by unbalanced data sets. The ultimate goal is to provide actionable information that can guide demographic research and focused interventions in mental health care.

2 Research Question

This study seeks to address the following research question:
RQ1 : How can advanced natural language processing techniques classify mental health-related user-generated content into relevant categories, identify specific stressors associated with these categories, and prioritize responses based on urgency levels?
By addressing this question, the study investigates the common stressors associated with mental health categories such as anxiety, depression, and stress, as well as the application of models such as zero-shot classification and MentalBERT for accurate sentiment analysis and the use of urgency levels to support targeted mental health interventions. This research seeks to improve mental health monitoring systems and create a scalable platform for real-world applications.

3 Dataset

The dataset used in this study is the Kaggle dataset "Sentiment Analysis for Mental Health", a compilation of mental health-related text data aggregated from multiple Kaggle sources:

- 3k Conversations Dataset for Chatbot
- Depression Reddit Cleaned
- Human Stress Prediction
- Predicting Anxiety in Mental Health Data
- Mental Health Dataset Bipolar
- Reddit Mental Health Data

- Students Anxiety and Depression Dataset
- Suicidal Mental Health Dataset
- Suicidal Tweet Detection Dataset

The dataset contains 51,074 unique posts and is annotated into six categories: Anxiety, Depression, Bipolar, Personality Disorder, Stress, and Normal. It contains three attributes:

- **id:** A unique identifier for each entry.
- **statement:** The textual content of the post.
- **status:** The annotated mental health category.

The *status* property serves as a categorization label to separate mental health into six preset classes, while the *statement* attribute is used to analyze the postings' content. As suicide risk detection is beyond the scope of this study, 10,652 posts annotated as "Suicidal" are excluded from the dataset. This preprocessing step results in a final dataset of 40,422 samples utilized for analysis and model development.

For evaluating the model, a test set was created using posts from the dataset *"solomonk/reddit_mental_health_posts,"* (Solomonk) as referenced in the paper *"Privacy Aware Question-Answering System for Online Mental Health Risk Assessment."*

4 Assumptions

This work assumes that the dataset represents real-world mental health talks, as well as that the classified categories accurately reflect users' mental health conditions. Posts with the category "Normal" are assumed to be free of mental health issues or hidden stresses. Furthermore, VADER mood evaluations are intended to effectively reflect the levels of urgency in textual content, and zero-shot classification predictions should closely match human judgments in distinguishing stressors.

5 Methodology

The methodology used in this study includes a number of crucial steps, including sentiment analysis, zero-shot classification, model fine-tuning, and data preprocessing. For user-generated content to be accurately classified and provide insightful information about mental health triggers (Chhikara et al., 2023), each step is essential.

5.1 Data Preprocessing

The training dataset is loaded into a pandas DataFrame, and entries labeled "Suicidal" are removed to focus on relevant mental health categories. Rows with missing *status* values are removed. SpaCy is subsequently used to process the text in the posts, which is transformed to lowercase, tokenized, and lemmatized to preserve base word forms. Stopwords and punctuation are removed to ensure meaningful content. The dataset is then divided into two sets: training (80%) and testing (20%), ensuring class balance by stratification. For optimal consistency with the classification model, labels are mapped to unique integer IDs using a dictionary. Finally, the training and testing data are converted into the Hugging Face Dataset format, which allows them to work easily with the Transformers library for tokenization and model training.

5.2 Fine-Tuning RoBERTa

The pretrained `roberta-large` model is loaded and configured for a classification task with six mental health categories. Training arguments are defined to optimize performance, including a small learning rate of 5e-5, gradient accumulation for an effective larger batch size, and mixed precision training with `fp16` for efficiency. Using Hugging Face's `Trainer`, the model is fine-tuned on the training dataset over two epochs, with evaluations conducted after each epoch. Accuracy and F1-score metrics are measured in the evaluation phase, where the best performance of the model is saved for deployment. The fine-tuned model is used to make predictions on the test data, demonstrating its ability to classify posts into predefined categories of mental health while leveraging the language understanding capabilities of the pretrained model. The fine-tuned model is saved as `fine_tuned_roberta_model`. The validation dataset `MentalHealth_TestSet.csv` is used to predict post labels, and the results are stored in `TestDatasetPredictions.csv`.

5.3 Stressor Detection Using Zero-Shot Learning

Zero-shot learning employs `roberta-large-mnli` model to categorize text into predefined categories without prior training in those categories. The goal is to classify the unlabelled posts into predefined categories, including *"Health Issues,"*

"Relationship Issues," "Financial Stress," "Workplace Stress," and *"Social Isolation."* The model is first initialized using Hugging Face's pipeline API, and the dataset is then preprocessed by eliminating rows labeled *"Normal,"* which are saved separately and prefilled with *"No Issues"* for both `ZeroShot_Cause` and `FineTuned_Cause`. The remaining dataset is converted to the Hugging Face Dataset format and examined in batches with a classification function that uses the `zero_shot_classifier` to select the most probable category for each sentence based on the highest confidence score. Following classification, the *"Normal"* rows are re-added to the dataset, and the full findings, including suggested causes, are saved to `zero_shot_predictions.csv` for later processing.

5.4 Fine-Tuning MentalBERT on Zero-Shot Learning Results

The MentalBERT model in the study (`mental/mental-bert-base-uncased`) is fine-tuned to classify textual data into specified stressor categories using zero-shot predictions as labeled input. The `zero_shot_predictions.csv` file is loaded, and the dataset is prepared by mapping stressor categories to numerical labels. Text data is tokenized with the MentalBERT tokenizer, truncated and padded, and stored as a Hugging Face Dataset object. The MentalBERT model is set up for sequence classification with six output labels: *"Health Issues," "Relationship Issues," "Financial Stress," "Workplace Stress,"* and *"Social Isolation"*. Training arguments are specified to optimize the process, such as a learning rate of 5e-5, gradient accumulation for an effective batch size of 16, and two training epochs. A Hugging Face Trainer is used to fine-tune the model, which is then saved locally for future use as `fine_tuned_mentalbert_cause_classifier`. For inference, the fine-tuned model is reloaded, moved to the appropriate device (GPU or CPU), and used to predict stressor categories for new text inputs via a function that tokenizes the input, processes it on the device, and outputs the predicted category based on the highest-scoring logits. Causes are then predicted using the validation dataset `MentalHealth_TestSet.csv`, and the results are saved in `test_data_with_predicted_cause.csv`.

5.5 Urgency Detection

The urgency of the posts is detected by assessing postings' sentiment with the VADER SentimentIntensityAnalyzer and assigning urgency degrees depending on the sentiment scores. It initially loads a dataset containing posts and projected labels, setting up VADER to compute sentiment ratings. To communicate a sense of urgency, each post is allotted a compound emotion score that is normalized to a 1–10 scale. Posts tagged as *"Normal"* have an urgency score of *"NA,"* suggesting a low level of urgency and no request for help. Urgency scores are used to classify posts as high (≥ 7), moderate ($4 \leq$ score < 7), or low (< 4), with high-urgency messages flagged for assistance. The enhanced dataset, which includes columns for urgency scores, levels, and flags, is stored to a new file named `test_data_with_urgency_flags.csv`, providing a scalable method for prioritizing posts that require quick attention.

6 Implementation Workflow

The implementation of this study involved the following sequential steps to classify mental health-related posts, identify stressors, and prioritize interventions:

Data Collection and Preprocessing

- Consolidation and cleansing of data from multiple Kaggle(Kaggle) sources are essential for data consistency.
- The removal of superfluous characters, normalization, and lemmatization all increase data quality for NLP applications.
- The exclusion of posts labeled "Suicidal" to balance the dataset is appropriate for the study's scope.

Model Training

- Fine-tuning the `roberta-large` model on a dataset for mental health classification.
- Using `roberta-large-mnli` for zero-shot classification to assign stressor categories to address the stressor labeling data gap.
- Fine-tuning MentalBERT (Yang et al., 2023) for stressor detection improves the system's precision.

Evaluation and Prediction

- Evaluating models using metrics such as accuracy and F1 scores ensures a quantitative assessment.
- Saving predictions for subsequent analysis is consistent with reproducibility and the requirement for additional insights.

Visualization

- Visualizations such as bar charts, histograms, and confusion matrices are generated for communicating results and identifying patterns.
- Frequencies, urgency levels, and categorization performances are highlighted.

Urgency Detection

- VADER sentiment analysis(Hutto and Gilbert, 2014) is used to assign urgency scores based on compound sentiment scores.
- Flagging posts for immediate attention ensures practical applicability in prioritizing interventions.

Output Generation

- Fine-tuned models are saved. Causes and scores are also saved in CSV files for reproducibility and further analysis demonstrates good data management practices.

7 Implementation Environment

The implementation environment provided the necessary computational resources and software frameworks to execute the study effectively. The details are as follows:

7.1 Dependencies

The following libraries and tools were used in the implementation: `pip`, `python`, `graphviz`, `nltk`, `gensim`, `vaderSentiment`, `numpy`, `pandas`, `peft`, `matplotlib`, `plotly`, `requests`, `rouge-score`, `scikit-learn`, `regex`, `datasets`, `huggingface-hub`, `accelerate`, `graphviz`, `evaluate`, `joblib`, `loralib`, `scipy`, `seaborn`, `torch`, `torchvision`, `tqdm`, `transformers`, and `notebook`.

7.2 Hardware Setup

- **GPU:** NVIDIA A100 (40GB) used for training and fine-tuning transformer models.
- **CPU:** Intel Xeon (16 cores) utilized for preprocessing and zero-shot classification.
- **RAM:** 64GB supporting large-scale data processing and model execution.

7.3 Software and Tools

- **Operating System:** Ubuntu 20.04 LTS.
- **Programming Language:** Python 3.10.10.
- **Frameworks and Libraries:** Hugging Face Transformers, SpaCy, Scikit-learn, Matplotlib, Plotly, VADER.

7.4 Data Sources

- Kaggle Mental Health Datasets.
- Reddit Mental Health Posts.

7.5 Technical Configuration

- **Batch Size and Learning Rate Optimization:** Ensures efficient model training.
- **Mixed-Precision Training (FP16):** Reduces memory overhead during fine-tuning.
- **Gradient Accumulation:** Simulates larger batch sizes without exceeding hardware limits.

8 Results

The fine-tuned RoBERTa model performs well in classification, achieving an accuracy of 87.32%, precision of 86.99%, recall of 87.32%, and F1 score of 87.07% on the test set. The confusion matrix (Figure 1) demonstrates the model's capacity to distinguish between six mental health categories: *Anxiety*, *Depression*, *Bipolar*, *Personality Disorder*, *Stress*, and *Normal*. The *Normal* category has the most accurately categorized examples, but categories like *Personality Disorder* and *Stress* are more prone to misclassification due to insufficient representation in the training data. Despite these issues, the overall performance demonstrates the model's capacity to categorize mental health-related posts effectively.

The distribution of predicted labels across the dataset (Figure 2) provides additional insight into

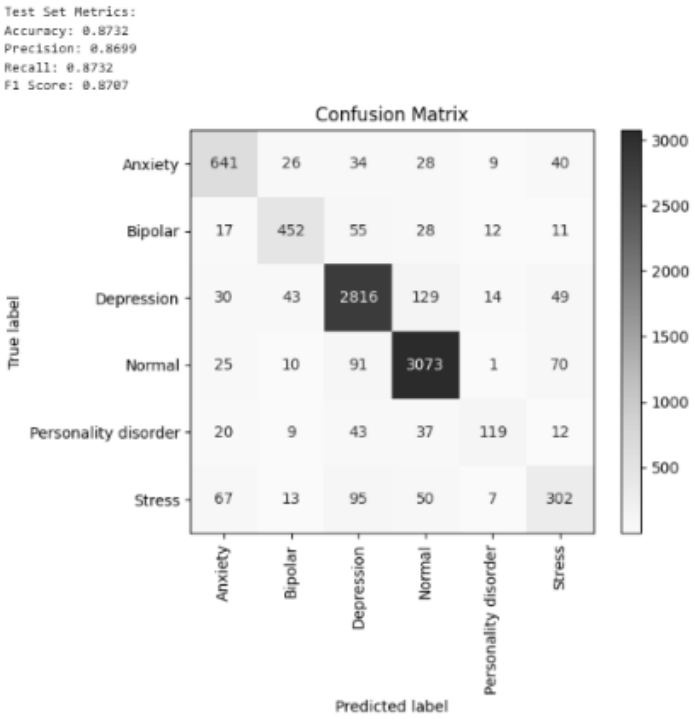

Figure 1: Confusion Matrix

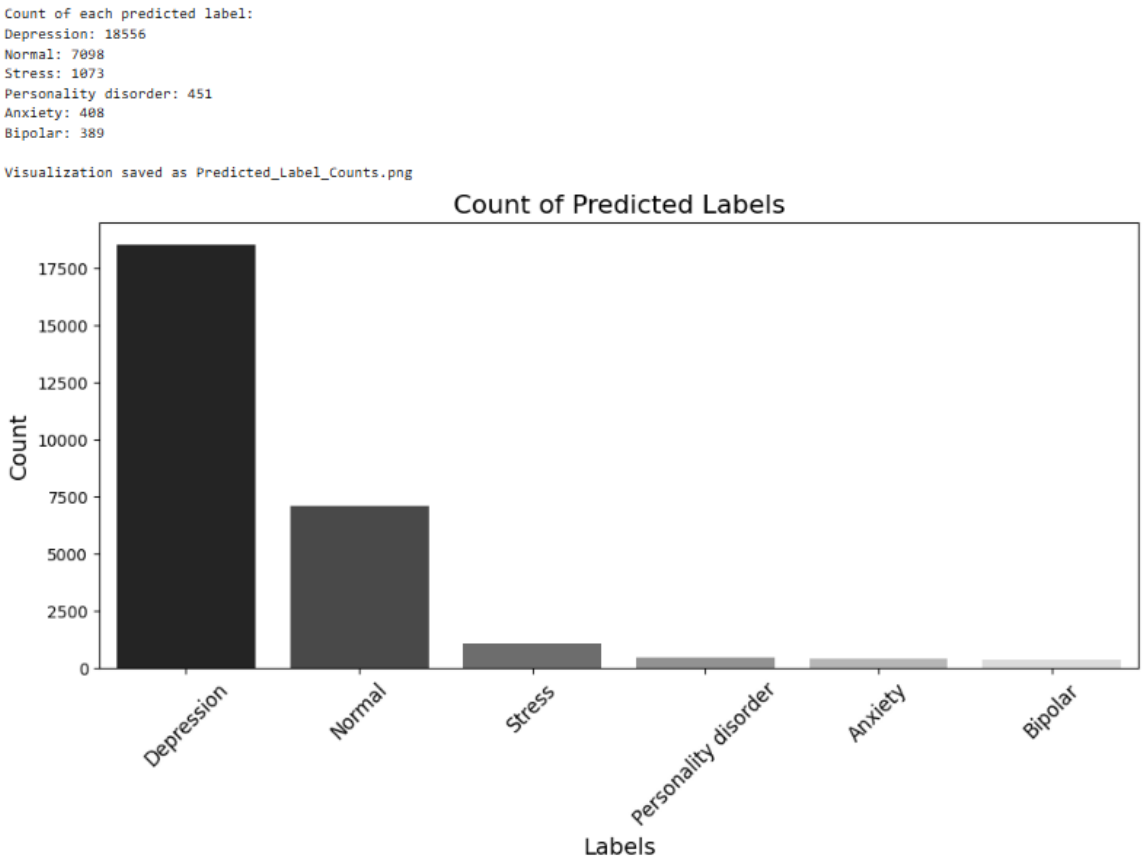

Figure 2: Label Distribution

the prevalence of mental health categories. *Depression* is the most commonly predicted classification, with 18,556 posts falling into this category, followed by *Normal* with 7,098 posts. Categories

```
Frequency of Stressor Types:
         Stressor Type  Count
0           No Issues   9740
1  Relationship Issues   7909
2     Social Isolation   5219
3        Health Issues   2799
4     Financial Stress   1845
5     Workplace Stress    463
```

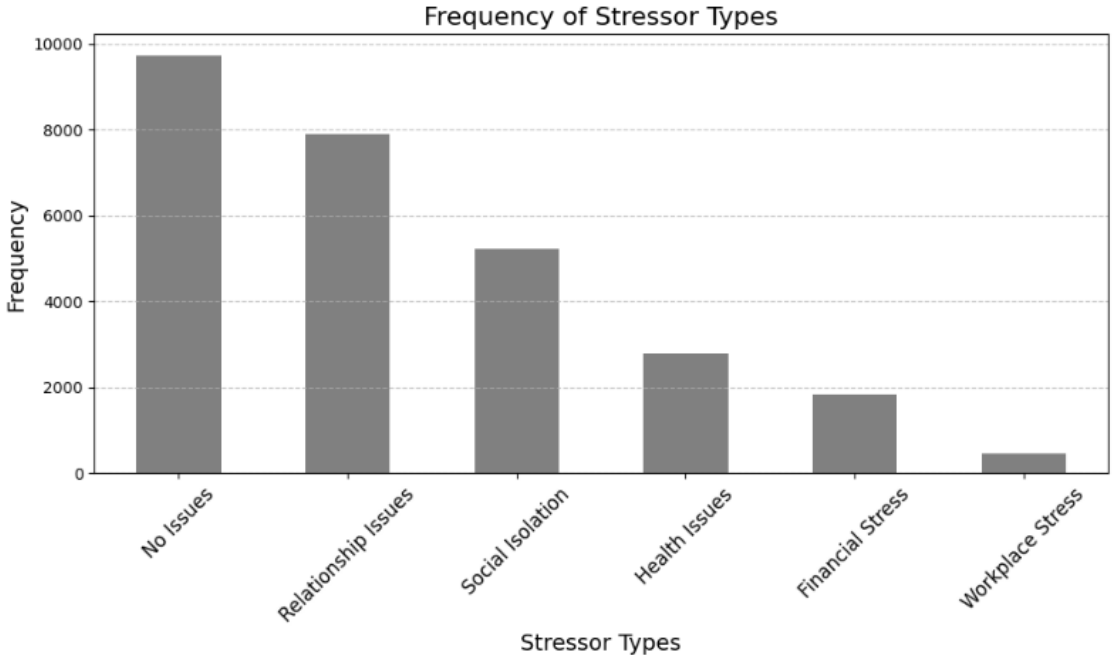

Figure 3: Stressor Categories

```
Counts of scenarios that require assistance and those that don't:
Flag for Assistance
No     21083
Yes     6892
Name: count, dtype: int64
```

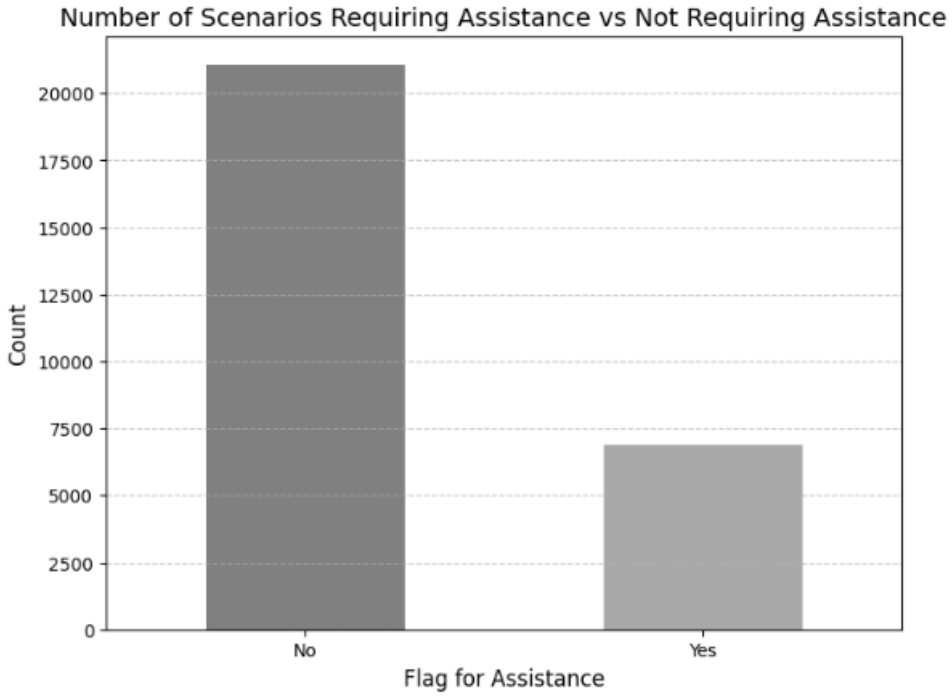

Figure 4: Urgency Distribution

like *Stress* (1,073 posts), *Personality Disorder* (451 posts), *Anxiety* (408 posts), and *Bipolar* (389 posts) are less frequent. This distribution illustrates the prevalence of depression-related topics in the analyzed content, consistent with current trends in mental health discourse.

The analysis of stressor categories, obtained from zero-shot classification and fine-tuned MentalBERT, identifies significant stressors affecting users. According to the bar chart (Figure 3), the most common stressor is *Relationship Issues*, accounting for 7,990 cases, followed by *Social Isolation* (5,219 instances) and *Health Issues* (2,799 instances). Posts classified as *No Issues* constitute the largest group, with 9,740 entries, while *Workplace Stress* has the fewest, with 463 entries. These results highlight the importance of interpersonal and societal elements as primary stressors in mental health issues.

Urgency detection using VADER sentiment analysis provides another layer of analysis by identifying posts that require immediate attention. Of the posts analyzed, 6,892 (24%) are flagged for immediate attention based on an urgency score greater than 7, with the majority of these posts relating to *Health Issues* and *Relationship Issues*. The remaining 21,883 posts (76%) are deemed lower priority. This prioritization paradigm underscores the importance of combining sentiment analysis and stressor detection to enable timely interventions. The urgency distribution, as shown in Figure 4, reveals that a considerable fraction of posts fall into the *Moderate Urgency* category, providing actionable insights for identifying posts with critical urgency levels.

9 Example Analysis and Model Evaluation

To evaluate the efficacy of the fine-tuned models and scoring mechanisms for stressor detection and urgency rating, the following examples are presented:

Example 1.

- ***Input Post:*** *"Strange. I don't have work today and I have a bit of free time so I can read shrill novels, but it feels weird. I've been nervous about checking the Google calendar just in case I read it wrong. But it's still weird, like you should be looking for a job to get rid of this feeling."*
- ***Predicted Mental Health Category:*** *Anxiety*
- ***Detected Stressor:*** *Workplace Stress*
- ***Urgency Score:*** *2.37*
- ***Urgency Level:*** *Low*
- ***Flag for Assistance:*** *No*

Example 2.

- ***Input Post:*** *"I don't like going out these days. No regrets or grudges/angry at things that have passed, and not worrying too much about the future—that's true serenity."*
- ***Predicted Mental Health Category:*** *Depression*
- ***Detected Stressor:*** *Social Isolation*
- ***Urgency Score:*** *7.95*
- ***Urgency Level:*** *High*
- ***Flag for Assistance:*** *Yes*

Example 3.

- ***Input Post:*** *"I am feeling great today. Everything is wonderful!"*
- ***Predicted Mental Health Category:*** *Normal*
- ***Detected Stressor:*** *No Issues*
- ***Urgency Score:*** *NA*
- ***Urgency Level:*** *Low*
- ***Flag for Assistance:*** *No*

10 Limitations and Challenges

Procuring mental health datasets was the first initial challenge faced in the study due to license issues. Determining the correct way (Li et al., 2023) to determine urgency involved a detailed literature domain review before implementations. Finetuning the pretrained models on a limited dataset involving a lot of exerimentation to finalize the parameters that delivered good results. Kaggle datasets were chosen instead of webscraping the internet for posts for these reasons. The study also has uneven datasets and the possibility of category bias due to underrepresented categories such as personality disorder. User-generated content lacks context, creating confusion and making it difficult to

identify implicit pressures. The ability to manage language and cultural oddities specific to mental health talks may be limited by reliance on pretrained models such as RoBERTa (Liu et al., 2019) (Devlin et al., 2018) and MentalBERT (Ji et al., 2021). Furthermore, sentiment-based urgency rankings may not always accurately reflect the severity of mental health disorders, and fine-tuning large transformer(Wolf et al., 2020) models requires significant computing resources, limiting experimentation.

11 Conclusion

This study illustrates the ability of advanced natural language processing approaches, including as fine-tuned transformer models and zero-shot classification, to assess mental health-related user-generated information. The approach provides practical recommendations for mental health interventions by quickly sorting postings into mental health categories and identifying stressors such as "Relationship Issues" and "Workplace Stress". The findings emphasize the frequency of depression and the need of urgency detection in prioritizing solutions for people who need immediate help. Despite obstacles such as class imbalance and dependency on pretrained models, the study's findings highlight the scalability of merging fine-tuned models with sentiment analysis for real-world use. Future study could look into how multimodal data like user behavior and ambient metadata, can be leveraged to increase stressor detection accuracy. Improving urgency detection models to include more subtle emotion markers could also assist target mental health interventions more effectively.

12 Acknowledgments

We are grateful to the contributors of the Kaggle and Hugging Face datasets for supplying the foundational data for this work. Appreciation is also extended to RIT for providing computational resources and useful guidance during the research.

References

P. Chhikara, U. Pasupulety, J. Marshall, D. Chaurasia, and S. Kumari. 2023. Privacy aware question-answering system for online mental health risk assessment. *arXiv preprint*.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint*.

C. J. Hutto and E. Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International Conference on Weblogs and Social Media (ICWSM-14)*.

S. Ji, T. Zhang, L. Ansari, J. Fu, P. Tiwari, and E. Cambria. 2021. Mentalbert: Publicly available pretrained language models for mental healthcare. *arXiv preprint*.

Kaggle. Sentiment analysis for mental health. Retrieved from `https://www.kaggle.com`.

A. Li, L. Ma, Y. Mei, H. He, S. Zhang, H. Qiu, and Z. Lan. 2023. Understanding client reactions in online mental health counseling. *arXiv preprint arXiv:2306.15334*.

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, and V. Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint*.

Solomonk. Reddit mental health posts dataset. Retrieved from `https://huggingface.co/datasets/solomonk/reddit_mental_health_posts`.

H. Suominen. 2014. Text mining and information analysis of health documents. *Artificial Intelligence in Medicine*, 61(3):127–130.

T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, and A. M. Rush. 2020. Transformers: State-of-the-art natural language processing. *arXiv preprint*.

K. Yang, S. Ji, T. Zhang, Q. Xie, Z. Kuang, and S. Ananiadou. 2023. Towards interpretable mental health analysis with large language models. *arXiv preprint arXiv:2304.03347*.