Copenhagen Business School

Data Science and Business Administration

Natural Language Processing and Text Analytics

- (CDSCO1002E) - Oral exam based on written product (EC) -

Authors:

Georgios Tzimas — 176995

Alessio Desideri — 176184

Nikolaos Tsatsampas — 172760

Examiner: Rajani Singh

Group: Fri-161558-31

Date: 29 May 2025

Number of words: 3,854

Characters (with spaces): 26,767

Pages: 13

Mental Health Condition Detection

Georgios Tzimas, Alessio Desideri, Nikolaos Tsatsampas
Copenhagen Business School
MSc. Business Administration and Data Science
{getz24ab, alde24ae, nits24ab}@student.cbs.dk

Abstract

In recent years, technological advancement has had a huge impact in various fields. Natural Language Processing (NLP) technologies are becoming more prevalent in healthcare, providing efficient and effective solutions to current problems. This study presents the development of NLP and text analytics techniques including Machine Learning (ML) and Deep Learning (DL) models for classifying text and identify mental health disease. Leveraging the kaggle dataset *"Sentiment Analysis for Mental Health"* (Sarkar, 2024), consisting of 51074 entries, we implemented classification algorithms as well as different embedding techniques (including Word2vec, BERT), and Large Language Models (LLMs). The experimental results demonstrate that Deep Learning models leveraging contextual embeddings from BERT and Large Language Models significantly outperform traditional Machine Learning algorithms and static embeddings like Word2Vec in accurately classifying mental health-related texts. Our best-performing model - Neural Networks with Bert Embedding - achieved an F1-score of 0.78, indicating robust predictive capabilities for early detection of mental health conditions through text analysis. These findings underscore the importance of advanced contextual embeddings in capturing nuanced linguistic patterns associated with mental health discourse. The study concludes that the implementation of NLP tools in clinical settings could support mental health screening and monitoring, contributing to more personalized and proactive care.

1 Introduction

"One in every eight people in the world live with a mental disorder" (Organization, 2022). Several conditions-including depression, anxiety, bipolarity, and others directly affect the well-being of individuals and the prosperity of societies in general. These conditions have direct consequences in how people think, feel and act, with their impact spreading from social and family relationships to government policies.

Despite medical breakthroughs in mitigating the symptoms, there are still problems that extend beyond medical treatments that need to be understood and addressed. A large proportion who is affected remain undiagnosed or untreated, due to *"the fear of what others might think, lack of insurance and unable to afford the costs of mental health care, lack of awareness that one is struggling with mental health and needs help"* (Turnbridge, n.d.) and more. Traditional diagnostic methods despite being well-used mechanism, they provide limitations that comes from inconsistencies, since they are based more on self-reported symptoms. For these reasons, creating mechanisms that can lead to timely and precise diagnosis is more than important.

The past few years Machine Learning and Artificial Intelligence techniques has uncovered new possibilities in healthcare sector. The application of these tools in mental diseases space presents a promising alternative to self-report assessment, but raises questions on which is the most effective technique to follow. Taking all the above facts into consideration, the following research question has been derived:

> *Which NLP approaches are most effective in identifying signs of mental disorders from conversational text?*

This paper address how NLP techniques can detect signs of mental distress that may not be clear. The main objective is the assessment of multiple NLP-based classification models-including traditional ML, NN, transformer-based models like BERT, and LLM (GPT-4o). This analysis contains insights into how different approaches interpet the text, comparing their results.

2 Related Work

Numerous recent studies on mental health classification have been deducted, even on the same dataset, leveraging ML and DL methods. This section highlights several pubicly available studies that contribute to this growing field of research.

First and foremost, Ding et al., 2025 in their study *"Trade-offs between machine learning and deep learning for mental illness detection on social media"*, they made use of the same dataset *"Sentiment Analysis for Mental Health"* (Sarkar, 2024). Their approaches consisted of Logistic Regression, Light Gradient Boosting Machine (LightGBM), as well as implementation of transformer-based architectures with focus on BERT and ALBERT. For binary classification, they restructured the original seven-class dataset into two categories: "Normal" (class 3) and "Abnormal" (merging the remaining six mental health-related classes). Their results showed that ALBERT significantly outperformed all other models, achieving an F1 score of 0.9650 and an AUC of 0.9928, while the best-performing traditional model, LightGBM, reached an F1 score of 0.9347 and an AUC of 0.9764. These metrics provide a strong benchmark against which other classification methods — such as those explored in our own study — can be compared.

The challenge of identifying multiple mental health conditions from social media, also addressed by Ameer et al., 2023. Their research, titled as *"Mental Illness Classification on Social Media Texts using Deep Learning and Transfer Learning"*, focuses on classifying five distinct mental disorders-depression, anxiety, bipolar disorder, ADHD, and PTSD-on user-generated content form Reddit. They also applied multiple models both from ML and DL. Among their comparison between Logistic Regression, Naive Bayes, SVM, CNN, RNN, and RoBERTa, they also ended up to the result that pre-trained on large corpora models, achieve better performance. More specific, RoBERTa reached an accuracy of 87.5% and F1-score of 0.85.

Similarly, Xu et al., 2024 explored hierarchical classification of mental health conditions on Twitter data using an ensemble approach combining traditional classifiers with contextual embeddings from DistilBERT. Their method not only addressed class imbalance but also aimed to maintain interpretability. The ensemble model achieved a macro F1 score of 0.88 across four categories and demonstrated improved robustness in detecting underrepresented conditions.

A comparison between our study and Ding et al., 2025 highlights key differences in both task setup and model performance. While our work tackled a multiclass classification problem across seven distinct mental health categories, Ding et al. simplified the task into a binary classification ("Normal" vs. "Abnormal"). However, despite this difference, both studies applied a progression from traditional machine learning models to transformer-based architectures. Analytical result comparisons are represented on Results 5.

Overall, all of these recent research demonstrate the strong capabilities of transformer-

based models and how they can be more accurate than traditional ML models.

3 Methodology

3.1 Dataset Description

The dataset utilized in this paper *"Sentiment Analysis for Mental Health"* retrieved from the Kaggle platform, which is originally compiled by Suchintika Sarkar. The data is sourced from diverse datasets within the same platform including Social Media/ Reddit/ Twitter posts, and more. It is compromised of 51,074 unique entries with three features, namely `Unnamed: 0` (unique_id which is displayed in this way), `statement` (textual data or post), and `status` (tagged mental health status of the statement). `status` is consisting of seven categories, Normal, Depression, Suicidal, Anxiety, Stress, Bi-Polar, and Personality Disorder, making it a good fit for implementing supervised learning's techniques. Its initial shape is 53043 rows × 3 columns, with memory usage: 1.2+ MB.

#	Column	Non-Null Count	Dtype
0	Unnamed: 0	53043 non-null	int64
1	statement	52681 non-null	object
2	status	53043 non-null	object

Table 1: Initial DataFrame Info

Unnamed: 0	statement	status
0	oh my gosh	Anxiety
1	trouble sleeping, confused mind, restless hear...	Anxiety
2	All wrong, back off dear, forward doubt. Stay ...	Anxiety
3	I've shifted my focus to something else but I'...	Anxiety
4	I'm restless and restless, it's been a month n...	Anxiety
...	...	...
53038	Nobody takes me seriously I've (24M) dealt wit...	Anxiety
53039	selfishness "I don't feel very good, it's lik...	Anxiety
53040	Is there any way to sleep better? I can't slee...	Anxiety
53041	Public speaking tips? Hi, all. I have to give ...	Anxiety
53042	I have really bad door anxiety! It's not about...	Anxiety

Table 2: Initial DataFrame First/Last Rows

3.2 EDA/Cleaning/Preprocessing

The data analysis process includes some important steps. Firstly, performing an EDA on initial data to take a quick view and understand them. This procedure included an inspection of our columns. On this stage we dropped the column with the unique id's

as it would not provide any important information for our analysis. To understand the balance of the dataset, a class distribution graph was created and each class' count (Fig. 1 & Table 3) displaying the frequency of each mental health disorder category present in the `status` column.

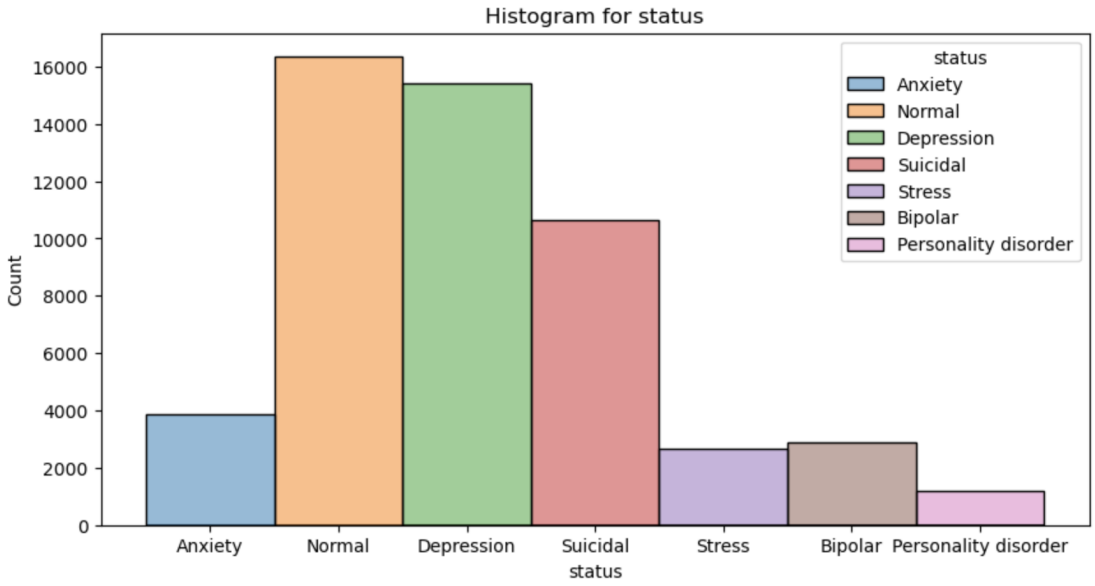

Figure 1: Histogram

status	**count (descending order)**
Normal	16351
Depression	15404
Suicidal	10653
Anxiety	3888
Bipolar	2877
Stress	2669
Personality disorder	1201

Table 3: Value Count of status

This is very useful for understanding data (im)balance. Later on section *" 3.3 Data Analytics: Modeling, Methods and Tools"*, we analyze how we dealt with it and on what score metrics we give more emphasis, while on section *" 5.2 Ethical Considerations"*, we speak about its ethical importance.

In addition, we plotted the most used words (Figure 2) to understand the dominant ones from all the texts.[1]

[1]Detailed word-cloud visualizations for each individual mental disorder class are provided in the appendix - Section A Explaratory Data Analysis (EDA) section

Figure 2: Most Common Words

Secondly, the cleaning process took place by calculating the number of null values. After identified that only 0.68% of the `statement`'s values are missing, they are simply dropped as this percentage is really low and it would have no significant difference to our analysis, and thus, no imputation needed. Moreover, duplicates were also dropped having already seen on EDA that there are proportionally low.

Thirdly, it is worth mentioning the early split on train-test set which was took place (80-20). This is a crucial step which help us to distinguish the train data in order to shuffle them and work on them without letting the test set to identify possible patterns of our dataset and avoid *"data snooping bias"* (Géron, 2019). The pre-process now start, by creating two functions which were applied both on train and test sets. The first one, called `clean`, for removing characters that are not necessary in a healthcare dataset, such as URLs, markdown-style links, all Twitter-style mentions that starts with "@", and punctuations. This function also lowercase the passed text. The second one, called `preprocess`, for tokenization and lemmatization. Lemmatization over stemming was chosen for normalizing the text, because our goal was to keep the tokens as accurate as possible, even if it is slower than stemming methods.

3.3 Data Analytics: Modeling, Methods and Tools

After the EDA, cleaning, and preprocessing phase, we proceeded with model development in combination with various text embedding techniques. Three traditional classification algorithms were implemented: Multinomial Naive Bayes, Logistic Regression, and Random Forest. In addition, a simple Neural Network was developed to test its performance with dense embeddings. We experimented with four embedding methods: CountVectorizer, CountVectorizer on Trigrams, TF-IDF, Word2Vec, and BERT. Initially, all tradi-

tional models were compared using CountVectorizer. For time constrain reasons, since Logistic Regression demonstrated better performance, we assumed that it is viable to keep, test and evaluate this model only with the all different embedding methods including Word2Vec and BERT embeddings. Furthermore, both Word2Vec and BERT embeddings were used with Neural Networks to compare performance across dense and contextual representations. Model effectiveness was assessed using confusion matrices—visualized as heatmaps—and classification reports on both the training and test sets. To address the class imbalance (during model training) we identified, we made use of the function `class_weight='balanced'`, which adjusts the weight of each class inversely proportional to its frequency. Additionally, the use of `StratifiedKFold` during cross-validation, preserved class distribution in each fold, important for dealing with over(/under)fitting. As far as the performance metrics are concerned, the evaluation of our models was based on several ones. Starting with precision, recall and F1-score, we based our comparisons more on F1 as it balances precision and recall as shown in the formula:

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{1}$$

Additionally, we examined the confusion matrices to gain insights into true and false predictions across classes in order to diagnose model errors.

Last but not least, a large language model (LLM), specifically GPT-4o, was employed as an advanced NLP approach to explore its capability in mental disorder classification from text. For this reason, we created a dataframe out of the shuffled, cleaned and preprocessed test set, consisting of 10,219 entries with their respective labels (Table 4) for feeding the GPT-4o.

#	**Column**	**Non-Null Count**	**Dtype**
0	text	10219 non-null	object
1	true_label	10219 non-null	object

Table 4: Test Set for LLM

Our approach followed through a progressive prompting strategy. We began with zero-shot learning, where the model was asked to predict the label from the text without prior examples. Next, we tried out a few-shot learning. We created another dataframe out of trained and preprocessed set of X with their corresponding labels of y. After this, we shuffled them to make sure it will generalize well enough, and provided two examples per class from the dataset, covering all seven mental health categories. This helped the model understand the structure of the task and align more closely the text with its respective label, as well as, in generating more accurate predictions.

The results of these comparisons are presented later in 4 Results section.

4 Results

This section presents the key results from the experimental evaluation of various machine learning and language modeling techniques for the multiclass classification of mental health-related texts. The evaluation followed a progressive structure, beginning with traditional machine learning models, transitioning through neural network architectures, and finishing with the use of large language models. Emphasis was placed on weighted and macro F1-scores due to the class imbalance observed (Fig. 1 & Table 3).

4.1 Traditional ML models

In the first phase, three traditional classifiers were evaluated using unigram-based CountVectorizer inputs:. Among Multinomial Naive Bayes, Random Forest, and Logistic Regression the last one consistently outperformed the others, achieving a weighted F1-score of 0.72 and a macro F1-score of around 0.64. These results, validated using stratified K-fold cross-validation, indicate strong generalization performance and minimal overfitting. At class level, Logistic Regression exhibited strong recall and precision for high-frequency categories such as "Normal", "Anxiety", and "Depression", while performance dropped considerably for underrepresented classes like "Stress" and "Personality Disorder." Based on its superior performance in this initial comparison, Logistic Regression was retained as the sole model for subsequent experiments focused on varying input representations.

label	precision	recall	f1-score	support
Anxiety	0.71	0.74	0.73	725
Bipolar	0.69	0.68	0.68	500
Depression	0.71	0.59	0.64	3019
Normal	0.87	0.94	0.91	3208
Personality disorder	0.41	0.44	0.43	179
Stress	0.42	0.51	0.46	459
Suicidal	0.62	0.65	0.64	2129
accuracy			0.72	10219
macro avg	0.63	0.65	0.64	10219
weighted avg	0.72	0.72	0.72	10219

Table 5: Logistic Regression with CountVect - kFold cross validation model performance

4.2 Input Representation Effects with Logistic Regression

The next series of experiments explored how different text vectorization techniques influence the performance of the most effective model, Logistic Regression. The representations tested included TF-IDF (unigrams), CountVectorizer with trigrams, TF-IDF

with trigrams, averaged Word2Vec embeddings, and fixed [CLS] embeddings from a pre-trained BERT model. Each of these representations provided a progressively richer and more context-sensitive encoding of the input text. TF-IDF representations produced results similar to CountVectorizer but showed slightly weaker recall on minority classes, resulting in a marginal drop in both weighted and macro F1-scores. The inclusion of trigram features via CountVectorizer and TF-IDF offered slightly better performance across all metrics, while when Logistic Regression trained both on averaged Word2Vec vector and BERT [CLS] embedding, performed poorer results, indicating that Logistic Regression is not powerful enough to leverage them fully.

4.3 Neural Networks

The third phase of evaluation examined whether non-linear modeling architectures could further improve classification outcomes. A two-layer feedforward neural network was applied to Word2Vec, while a deep Transformer-based architecture with 12 layers was applied with BERT embedding. When trained on Word2Vec vectors, the network achieved an accuracy of 0.7004 and a weighted F1-score of 0.6925. NN with BERT embeddings yielded further improvements especially through the fine-tuned BERT, with the model achieving a test accuracy of 0.7816 and a weighted F1-score of 0.7810. Most notably, it produced significantly improved recall scores for underrepresented classes such as "Stress" and "Personality Disorder," more than doubling the recall observed under the baseline logistic regression model.

label	**precision**	**recall**	**f1-score**	**support**
Anxiety	0.83	0.83	0.83	725
Bipolar	0.81	0.75	0.78	500
Depression	0.72	0.73	0.73	3019
Normal	0.92	0.93	0.93	3208
Personality disorder	0.72	0.50	0.59	179
Stress	0.61	0.59	0.60	459
Suicidal	0.67	0.69	0.68	2129
accuracy			0.78	10219
macro avg	0.76	0.72	0.73	10219
weighted avg	0.78	0.78	0.78	10219

Table 6: BERT Classification Report

4.4 LLM - GPT-4o

The final set of experiments evaluated GPT-4o, OpenAI's flagship large language model, using zero-shot and few-shot classification. The prompt that was given both on zero and

on few-shot, is the following:

```
"You are a classifier.  You **must** choose exactly one label from this
list and reply" "with **only** that label (no extra words)"
```

On zero-shot prompting technique, when evaluated on a 1,000-sample subset, achieved a weighted F1-score of 0.65, a macro F1 of 0.55. Scaling to the full 10,201-sample test set, results remained consistent, with a weighted F1 of 0.65 and macro F1 of 0.55.

The few-shot configuration on 1,000-sample test set presented improvements compared to zero-shot with a weighted F1-score of 0.67, macro F1 of 0.5. When extended to the full 10,187-sample dataset, few-shot prompting achieved even greater gains. With weighted F1 of 0.69, macro F1 of 0.60, it approached the performance of neural network models trained on BERT embeddings.

Although both prompting strategies performed well, few-shot prompting clearly improved class-level performance, particularly on high-frequency categories. However, performance remained lower for minority classes such as "Stress" and "Personality Disorder," where recall continued to lag.

4.5 Actionable Insights

The findings of this study reveal important insights into how different machine learning methods perform in classifying mental health-related text. The progression from traditional models to neural networks and large language models not only demonstrates increasing performance but also signals the evolving trade-offs between simplicity, generalization, and contextual understanding.

Logistic Regression proved to be the strongest among the traditional models and it was a good baseline for testing other text representations. Subsequent experiments revealed that changing text representations significantly affected performance. Trigram-based vectorizers improved both macro and weighted F1 scores compared to unigrams. However, the performance declined a bit, when dense semantic embeddings like Word2Vec and BERT [CLS] were used with Logistic Regression, indicating that the model struggled to exploit the richer representations and that the more expressive models were needed to fully leverage these embeddings.

Neural networks pushed performance higher, particularly when trained on BERT embeddings. The best results came from a fine tuned BERT - deep transformer with 12+ layers - which reached the highest accuracy and F1 scores in the whole study. More importantly, it also significantly improved the predictions for the minority classes, showing its strength in learning subtle patterns across all categories.

GPT-4o, tested in zero-shot and few-shot configurations, showed that large language models can be effective without any fine-tuning. While zero-shot results were solid, few-shot prompting (with just two examples per class) led to significant performance gains,

but these advancements came at a substantial cost (x8 increace). However, GPT-4o also struggled with the same low-frequency classes, suggesting that even advanced LLMs have limits when not further fine-tuned.

Overall, the study reveals a clear performance hierarchy: fine-tuned neural architectures with pretrained embeddings deliver the highest accuracy and fairness across categories, few-shot prompting with LLMs provides a compelling alternative when training is not feasible; and traditional models remain competitive when paired with optimized vectorizers. Future model selection should balance predictive performance with resource constraints such as data volume, computational costs, and ease of deployment.

4.6 Valuable Outcomes

The results exhibit significant applicability in a variety of sectors of healthcare ecosystem, such us clinics and digital health applications. All of the methods used have very similar results, demonstrating strong predictive performance in detecting mental health conditions from text.

First of all, these models can be integrated into healthcare clinics' systems to analyze patient-reported text data and assist clinicians on early and accurate detection of mental health, prioritizing individuals at risk of self-harm.

Similarly, mobile applications that offer mental health support could utilize such models that can be coupled with voice-to-text applications. Guided prompts or questions asked by machines to users, along with their voice recording that is transcribed into text, can be analyzed and enable continuous and passive mental health screening. This hybrid approach enhances accessibility for users: (a) who can not afford a clinical diagnosis, and (b) who may find it more comfortable to communicate verbally. In this way, with just a mobile phone, everyone can have access to an atomic detection of their mental condition from the comfort of their home.

The study also highlights that options that are both time and cost efficient can be viable options for deployment at scale. While fine-tuned BERT offers the best performance, it requires significant time, compute resources, and technical know-how.

5 Discussion

To contextualize our model performance, we compare our findings with those reported by Ding et al. (2025), who addressed a related mental health classification task. This comparison highlights how differences in task design and modeling choices influence outcomes. In our evaluation, Logistic Regression on BERT embeddings achieved a strong weighted F1-score of 0.74, while fine-tuned BERT on Neural Networks reached 0.778. Large Language Models like GPT-4o, though competitive in a few-shot setup (weighted

F1 = 0.69), still lagged behind supervised neural models in performance, particularly for minority classes. By contrast, Ding et al. reported exceptional results with ALBERT, which achieved a binary F1-score of 0.9650 and AUC of 0.9928, outperforming traditional models like LightGBM. These findings underscore the impact of task complexity, dataset restructuring, and fine-tuning strategies on model effectiveness.

5.1 Limitations

While this study provides meaningful insights into the classification of mental health conditions using machine learning and language models, it is important to acknowledge some limitations that are presented.

The NLP models used in this study rely purely on textual content. However, social media posts often derive meaning from surrounding context such as post history or user metadata. Without this context, even the best models can misclassify emotionally charged statements. The dataset used provide also some strong limitations. While large and somewhat diverse, is still limited to a specific subset of platforms and users. As a result it likely overrepresents certain demographics (e.g., younger, English-speaking users) and underrepresents others. That implies that models trained on it may not generalize well to populations with different linguistic, cultural, or clinical expressions of mental illness. Lastly, the class labels used for training are based on annotations or self-reports, not verified clinical diagnoses. This introduces uncertainty around whether the labels truly represent diagnosable conditions, which may in turn affect the validity of the model's predictions.

An additional limitation of this study lies in the computational and financial constraints encountered during model training and evaluation. Fine-tuning the neural network with BERT embeddings and training Logistic Regression with trigram features at full scale was limited by the restricted GPU availability on Google Colab, which would require a paid subscription for optimal performance. Similarly, testing GPT-4o in both zero-shot and few-shot settings incurred a cost of approximately €30, limiting our attempt of experimentation and hyperparameter tuning that could be performed.

5.2 Ethical Considerations

Despite ML and DL advancements, several challenges remain, including ethics. A critical issue worth mentioning is possible bias in training data. Text, especially those who were gathered from social media may contain humor/sarcasm/irony which pure classification models cannot detect. Transformer-based models like BERT or LLMs can handle sarcasm at some point, but they are not perfect especially if sarcasm wasn't well-represented in the training data and this can be problematic in sensitive domains like mental-health leading to false positives/negatives.

Another important concern we recognize, especially working with sensitive mental health data is transparency, data collection and handling. While the dataset is publicly available and its author provide transparency regarding the combined data sources that used, it is not clear whether adequate user consent (European Union, 2016b), data minimization, purpose limitation and privacy protection (European Union, 2016a) were fully implemented at source level, even if they are anonymized according to GDPR.

Speaking about the models' credibility, we need to refer our concern on transparency and interpretability of the models themselves. Algorithms like Neural Networks and ensemble methods like Random Forest, are categorized as black-box models (SEON, n.d.). This term refers to models that once trained, provide limited insight into their internal decision-making process. This is due to their complexity (large number of trees, layerd architecture of NN), which do not let us identify how they make predictions or detect embedded biases.

6 Conclusion & Future Work

This study was conducted to investigate which NLP approach is more effective on identifying signs of mental disorders from text. Utilizing a dataset containing around 50,000 records, we implemented different embedding techniques on a specific model (LR) for identifying the most accurate, comparing it also with Neural Network outcomes and advanced Deep Learning models like BERT and GPT. The results demonstrate that all the aforementioned techniques are capable of classifying text to its according category on a satisfactory level of accuracy and robustness, particularly in the context of a real-world comment dataset. Overall, this paper integrates technical implementations with thoughtful evaluation and ethical considerations, providing societal value for early detection and prevention of mental health issues and emotional distress that can lead to life-threatening situations.

One part of our next steps involves experimenting with additional ML models like K-Nearest Neighbors (KNN), Support Vector Machines (SVM), or other ensemble methods rather than Random Forest, such as XGBoost, suitable for classification tasks. At the same time, we plan to test them with different embedding techniques as we previously did with Logistic Regression. We also aim to optimize/fine tune the best-performing baseline model, starting with Randomized Search (for a faster approximation of promising parameter values), followed by GridSearch for more precise fine-tuning (most effective but requires higher computational cost).

Another promising plan is to test the same predictive model on more datasets - the merrier, the better. This will impact our assessment of the model's generalizability and robustness across different contexts. Furthermore, this will reveal potential weaknesses that may arise from vocabulary differences etc.

References

Ameer, I., Arif, M., Sidorov, G., G´omez-Adorno, H., & Gelbukh, A. (2023). Mental illness classification on social media texts using deep learning and transfer learning. *arXiv*. %5Curl%7Bhttps://arxiv.org/abs/2207.01012%7D

Ding, Z., Wang, Z., Zhang, Y., Cao, Y., Liu, Y., Shen, X., Tian, Y., & Dai, J. (2025). Trade-offs between machine learning and deep learning for mental illness detection on social media. *Scientific Reports*, *15*. https://doi.org/10.1038/s41598-025-99167-6

European Union. (2016a). Regulation (EU) 2016/679 (General Data Protection Regulation), Article 5: Principles relating to processing of personal data.

European Union. (2016b). Regulation (EU) 2016/679 (General Data Protection Regulation), Article 7: Conditions for consent.

Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow* (2nd). O'Reilly Media.

Organization, W. H. (2022). *Mental disorders* [https://www.who.int/news-room/fact-sheets/detail/mental-disorders].

Sarkar, S. (2024). *Sentiment analysis for mental health* [https://www.kaggle.com/datasets/suchintikasarkar/sentiment-analysis-for-mental-health/data].

SEON. (n.d.). *Black box machine learning*. https://seon.io/resources/dictionary/blackbox-machine-learning/

Turnbridge. (n.d.). *Undiagnosed mental illness: What you should know* [https://www.turnbridge.com/news-events/latest-articles/untreated-undiagnosed-mental-illness/].

Xu, F., et al. (2024). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2401.04655*. https://arxiv.org/abs/2401.04655

Appendix

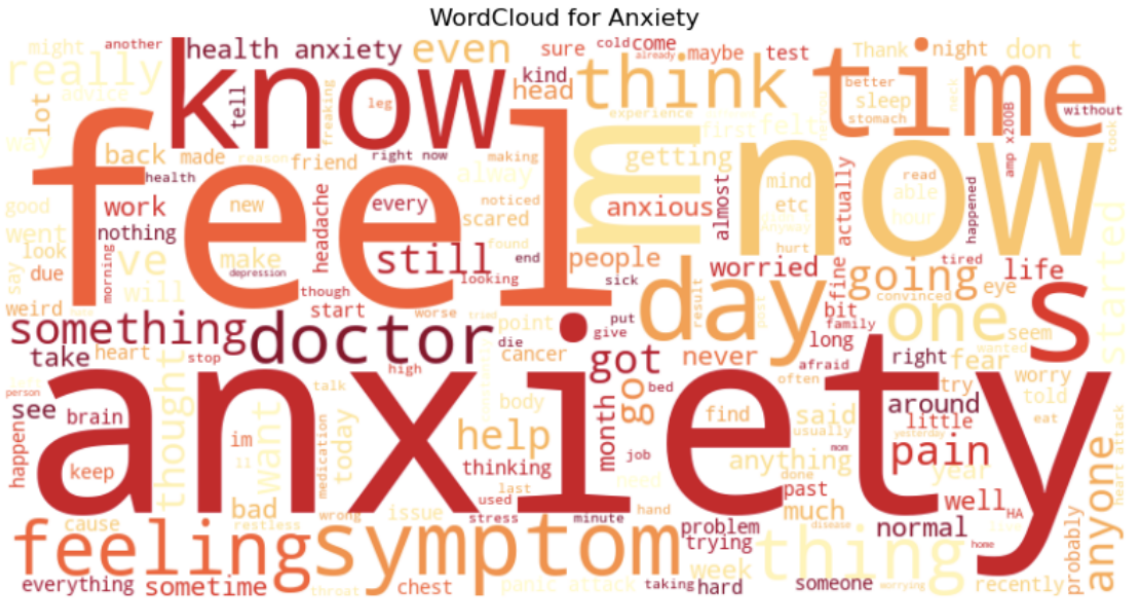

Figure 3: Most Common "Anxiety" Words

Figure 4: Most Common "Bipolar" Words

Figure 5: Most Common "Depression" Words

Figure 6: Most Common "Normal" Words

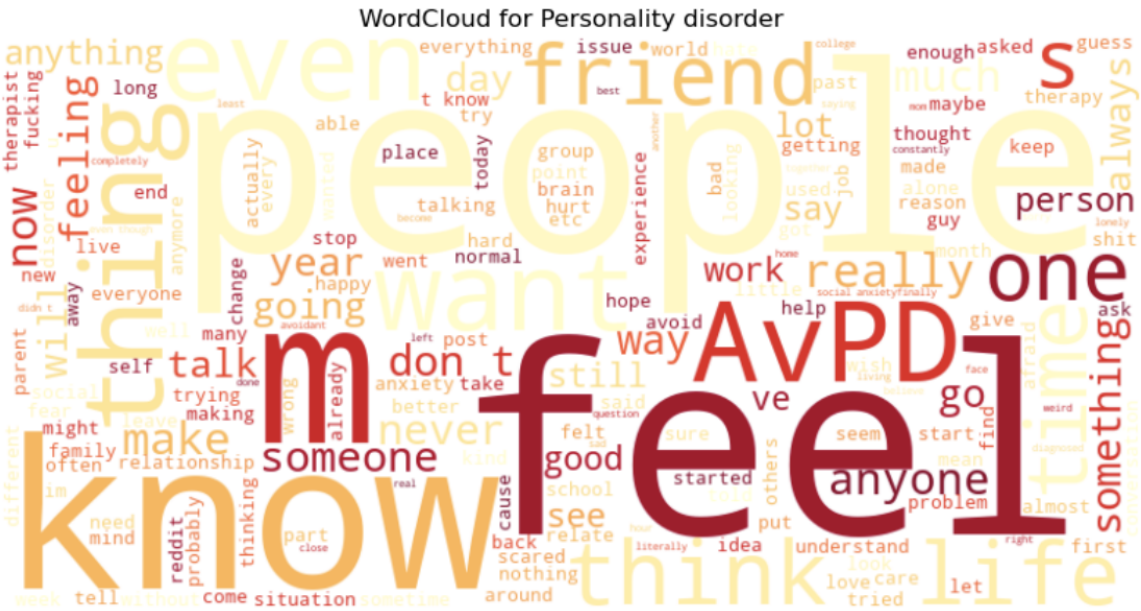

Figure 7: Most Common "Personality Disorder" Words

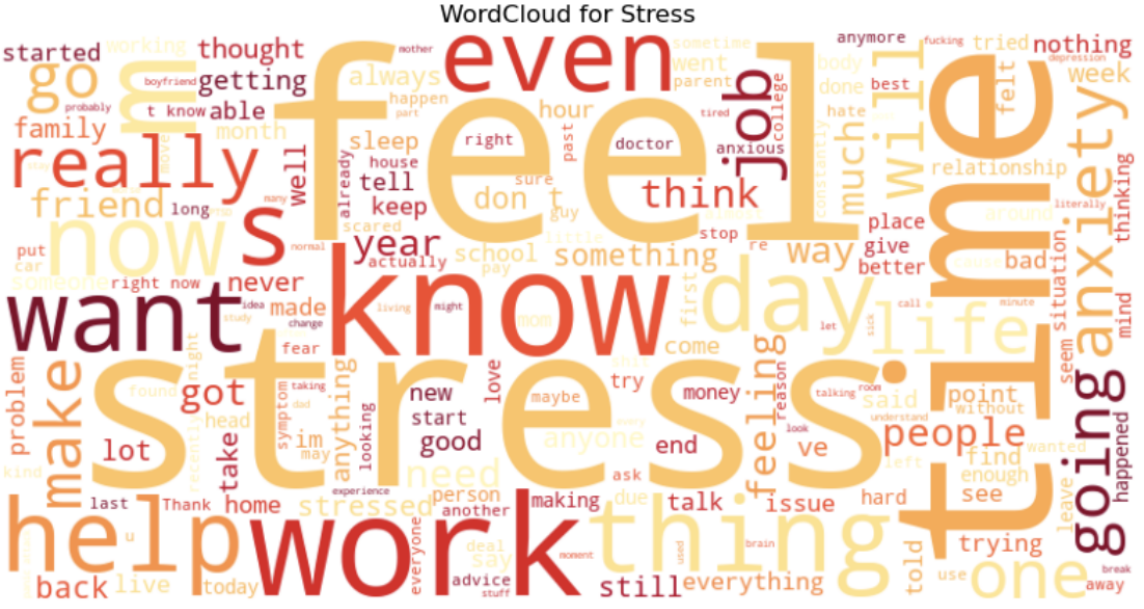

Figure 8: Most Common "Stress" Words

Figure 9: Most Common "suicidal" Words

Figure 10: Naive Bayes - Confusion Matrix (Test Data)

Figure 11: Logistic Regression - Consfusion Matrix (Test Data)

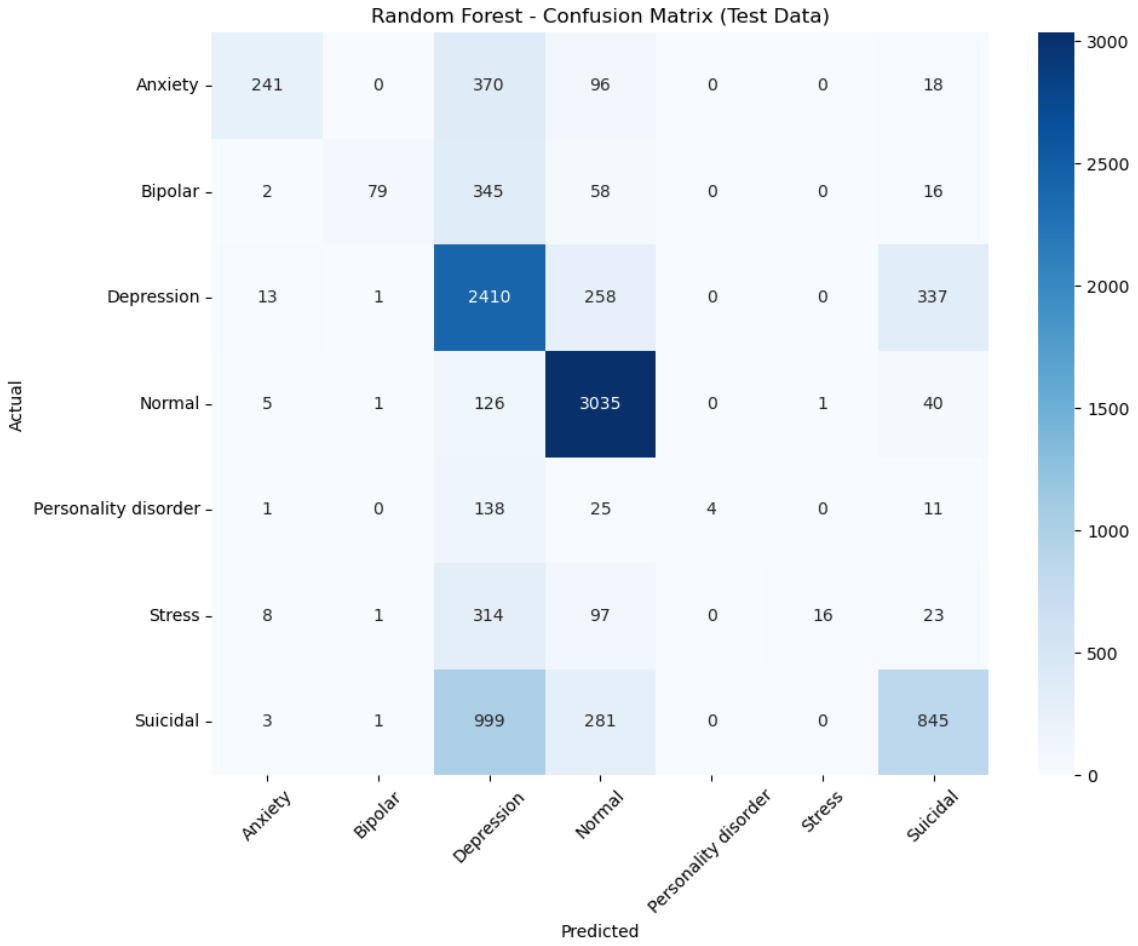

Figure 12: Random Forest - Consfusion Matrix (Test Data)

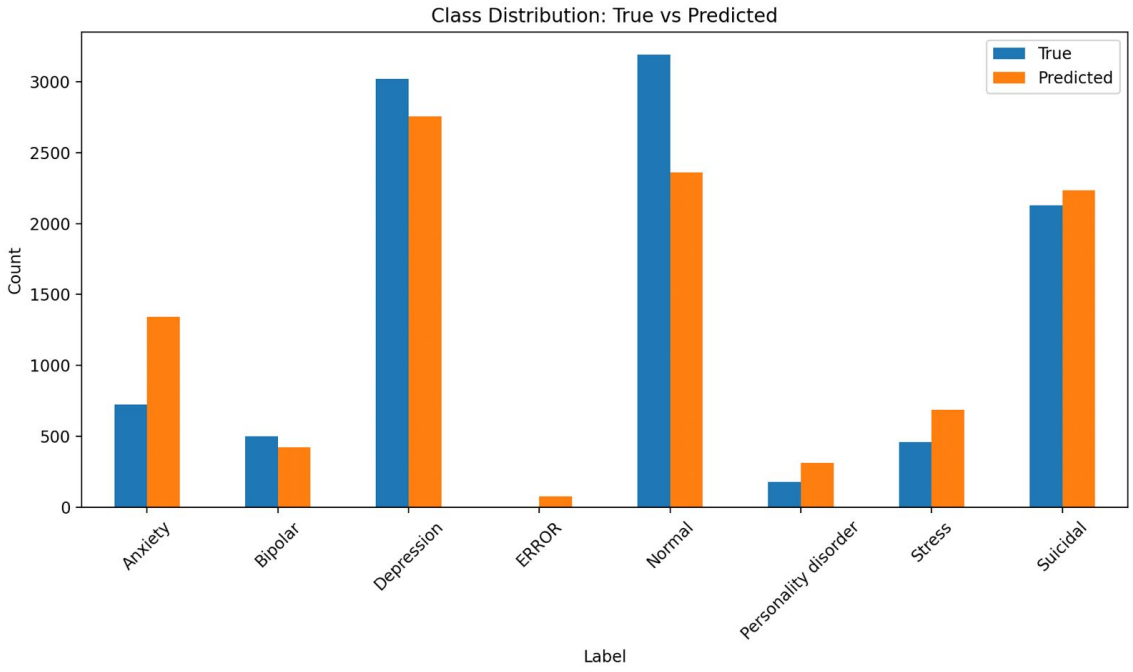

Figure 13: Class Distribution (Full Test Set) - Zero Shot

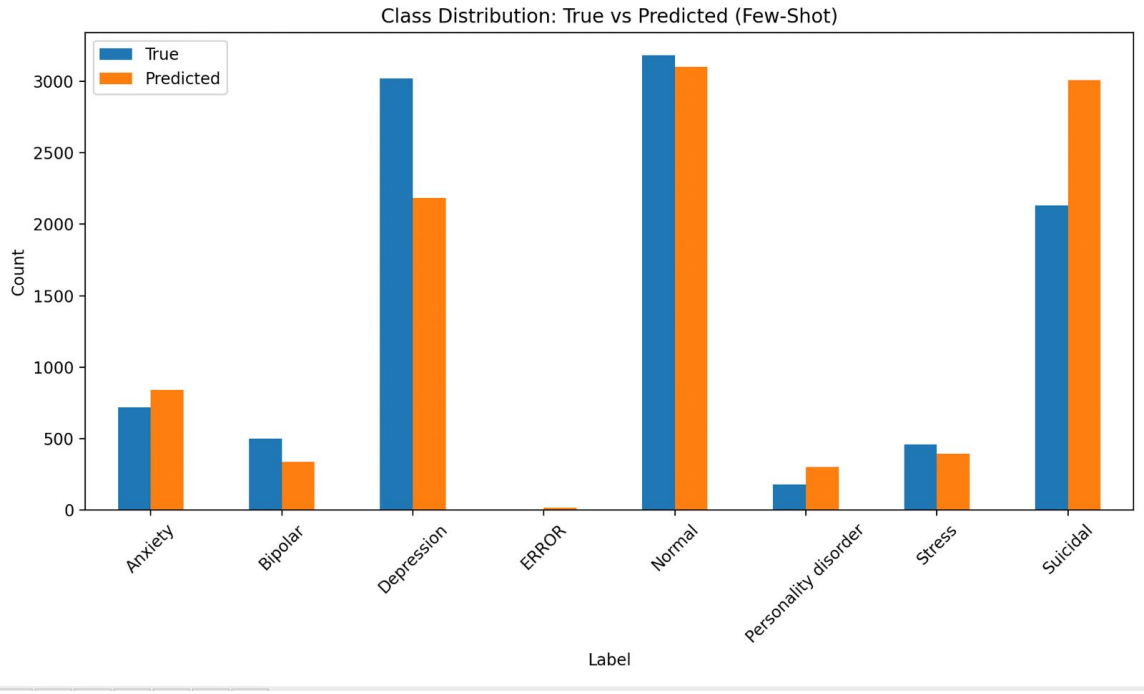

Figure 14: Class Distribution (Full Test Set) - Few Shot

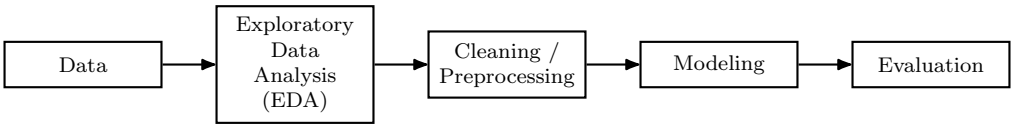

Figure 15: Conceptual Framework Diagram