Detection and classification of mental illness based on social media posts

Muthumayan Madhayyan, Bhaskar Borah

Summer 2024

Abstract

WHO estimates that 1 in 8 people live with a mental disorder. However, mental health does not get the same importance as other public health issues (say a viral infection). Part of the problem is in obtaining real-world, real-time demographic data on people impacted and the type/frequency of mental health afflictions that they deal with. Social media activity (if monitored ethically) can provide informal quantitative data on types and demographics of mental health afflictions. These are not clinical evaluations but serve as anonymized approximations of underlying conditions. This is where NLP comes in. We intend to train a RoBERTa based classifier that can then work on unseen social media posts and classify them. To this end, we plan to train the classifier using datasets obtained from mental-health subreddit channels. The choice of RoBERTa is based on data from past researchers.

1 Introduction

Despite the crucial role mental health plays in our well-being, it is often neglected and rarely discussed with the same seriousness as physical health. Some social media platforms, such as Reddit, have created online forums dedicated to mental health issues. These forums act as informal support groups for individuals seeking help with their mental health conditions.

We aim to fine-tune a pre-trained language model with social media data to create a classifier that accurately categorizes posts into mental health illness categories. This model would identify mental health trends across various demographics and events but is not intended to replace formal diagnostics due to privacy concerns and the lack of professional evaluation.

The motivation behind developing this classifier is to enhance public mental health monitoring in a non-intrusive manner. This can help public health services make informed decisions on resource allocation, and design targeted interventions ultimately improving community well-being.

Murarka et al. (2021) classified five major mental illnesses—depression, anxiety, bipolar disorder, ADHD, and PTSD—using unstructured Reddit data and created datasets with posts, titles, and both. While prior research has focused on classification with single or multiple data sources, there has been no direct comparison of mixed versus single-source data or evaluation of transformer models like GPT-2. We fine-tuned various pre-trained models, including GPT-2, DistilBERT, and XLM-RoBERTa, to assess their effectiveness for our classification task.

Comparing the effectiveness of various language models can lead to methodological improvements in mental health classification tasks. Early detection and intervention can reduce the long-term costs associated with untreated mental health conditions.

2 Background

Several research teams have attempted to classify social media posts ascribed to different types of mental health afflictions using either subreddit posts related to mental health or Twitter tweets related to depression.

Murarka et al. (2021)used LSTM, BERT, and RoBERTa models to classify mental illnesses into five categories, with RoBERTa yielding the best results through masking and synonym replacement techniques. Nadeem et al. (2022) achieved over 97% F1 scores for binary classification using a combination of CNN, LSTM, GRU, and self-attention but saw an 11 percentage point drop in F1 scores for a three-class multi-class classification. Kim et al. (2020) employed CNN-based and XG-Boost models to classify mental health conditions from Reddit posts, achieving strong F1 scores in binary classification. Aldhyani et al. (2022) used CNN–BiLSTM and XGBoost models for detecting suicidal ideation from Reddit posts, reaching 95% and 91.5% accuracy in binary classification. While each team achieved notable success with various ML algorithms, there were no comparative results on standardized datasets (such as Kaggle), but most projects attained F1 scores above 0.7.

We chose to build on the work of Murarka et al. due to their thorough data collection and cleaning process, well-balanced dataset, and detailed experimentation with stress tests. By replicating their experiments and using RoBERTa as a baseline, we aim to fine-tune various pre-trained language models like GPT-2, DistilBERT, and XLM-Roberta for our classification task. Our focus will remain on transformer-based architectures and enhancing explainability for reported misclassifications.

3 Methods

Dataset

Murarka et al. (2021) collected data (titles and posts) across subreddits representing each disorder category and a ‘none’ category, resulting in 13,727 posts for training, 1,488 posts for validation, and 1,716 posts for testing. These datasets have been utilized in our experiments.

Challenges in using the dataset

Classifying mental health from user posts presents several challenges, including:
Dataset Issues: Social media posts in subreddits like r/depression and r/anxiety are inherently noisy and lack precise definitions of mental illnesses due to symptom overlap and varied posting styles. The spectrum of mental illness (multi-labeling) is very likely in these datasets. Also, the emotional tenor of a post does not always directly relate to a specific mental health condition.
Training Issues: Sequence length and batch size limitations caused out-of-memory (OOM) errors, overfitting, long training durations, and errors in saving/loading fine-tuned models.
Inference and Analysis Issues: Explainability of mismatches.

Our Approach

Our experiments are by no means novel. They build on established transformer architectures and pre-trained models. We are attempting to improve outcomes by analyzing classification failures.

We inherit training/validation datasets used by past researchers and conduct different experiments listed below:
Experiment 1: We conducted two experiment variants: one using only the post body and the other using both the body and title. Past research suggests that titles improve classification accuracy.
Experiment 2: Multiple pre-trained language models are based on transformer architectures, and we experimented with a few - RoBERTa, Distilbert, XLM-RoBERTa, GPT2, etc.
Experiment 3: We experimented on the models with different hyperparameters
Experiment 4: Comparing social media posts with clinical definitions
Experiment 5: To understand the prediction data mismatches, we conducted experiments with different masking approaches to investigate the impact of class-name strings within the posts.

To evaluate our multi-class classification problem, we use traditional metrics like precision, recall, and F1 scores to maximize F1 scores.

4 Results and Discussion

4.1 Comparison Testing

To investigate and establish new baselines using the same architecture as Murarka et al.'s models, we conducted a comparison test to replicate their original models (noting that Murarka et al.'s code is not publicly available)

Architecture and Design

Murarka et al. (2021) fine-tuned BERT and RoBERTa base models with the Reddit dataset, utilizing pre-trained language models. They set a sequence length of 512, applying padding and truncation as required, and used a learning rate of 0.00001 with an Adam Optimizer. The model included a dropout layer with a rate of 0.3, ran for 10 epochs, and used a batch size of 32.

Training set	models	posts		titles		posts+titles	
		F1	Accuracy	F1	Accuracy	F1	Accuracy
Murarka et al.	BERT	0.82	0.82	0.71	0.71	0.87	0.87
	RoBERTa	0.86	0.86	0.72	0.72	0.89	0.89
Our Experiment	BERT	0.80	0.80	0.69	0.70	0.85	0.85
	RoBERTa	0.83	0.84	0.71	0.72	0.87	0.87

Table 1: Comparison test results per model

Results and Discussion

The comparison test results confirmed that Murarka et al. 's (2021) original experiments can be successfully replicated, validating the language model and hyperparameters they used.

4.2 Baseline Model

We selected the RoBERTa test results from Murarka et al.'s paper as our baseline because it was their most successful outcome among all experiments. We specifically chose data that includes posts plus titles. Our experiment with RoBERTa did yield results closer to those reported by Murarka et al.

experiment	posts		titles		posts+titles	
	F1	Accuracy	F1	Accuracy	F1	Accuracy
RoBERTa	0.86	0.86	0.72	0.72	0.89	0.89

Table 2: RoBERTa Model - Murarka et al.

4.3 Experiments 1 and 2

We combined experiments 1 and 2 into one step. This experiment aimed to investigate if executing a different set of models could improve the baseline. We ran all of these models for posts, titles as well as posts plus titles.

- DistilBERT is efficient and retains much of BERT's performance. We wanted to investigate how well it performs on this corpus of data.
- GPT-2's generative capabilities and large pre-training corpus enable it to handle complex language patterns and nuances effectively.
- XLM-RoBERTa, as it is trained on a bigger and more diverse dataset than Roberta (although the dataset being multilingual would not be much help in our case).

models	posts		titles		posts+titles	
	F1	Accuracy	F1	Accuracy	F1	Accuracy
DistilBERT	0.80	0.80	0.69	0.69	0.85	0.85
GPT2	0.79	0.80	0.68	0.69	0.84	0.84
XLM-RoBERTa	0.80	0.80	0.68	0.70	0.85	0.85

Table 3: Experiment 1 and 2 results

Results and Discussion

The results from these experiments are not an improvement over the RoBERTa baseline. This suggests that the RoBERTa model provides the best results using the original hyperparameters.

4.4 Experiment 3

Architecture and Design

We experimented with various hyperparameter combinations, finding that some significantly degraded performance while others yielded respectable results, using a sequence and embedding length of 512 for posts and posts plus titles, a learning rate of 0.00002 with the Adam optimizer, a dropout rate of 0.2, training for 12-15 epochs (for GPT batch size as 20 and epochs as 20), and a batch size of 30.

models	posts		titles		posts+titles	
	F1	Accuracy	F1	Accuracy	F1	Accuracy
DistilBERT	0.76	0.76	0.61	0.62	0.81	0.82
GPT2	0.74	0.77	0.64	0.66	0.82	0.83
XLM-RoBERTa	0.74	0.75	0.74	0.71	0.83	0.87

Table 4: Experiment 3 results

Results and Discussion

As seen in Table 4, we were not able to exceed the baseline model although we managed to get nearly similar results using XLM-RoBERTa.

Once our experiments were done, we wanted to analyze what could cause the mismatch in predictions. We started with an EDA of the training and test dataset and performed a detailed analysis of the data which we explore in the next sections.

4.5 EDA

For brevity, we present the salient observations during EDA. Details can be found in the Appendix.

1) The training, validation, and test datasets are well-balanced.
2) The class names appear in the posts/titles at a good proportion. Class name words like *adhd*, *bipolar*, and *ptsd* are concentrated in their respective categories, while *depression* and *anxiety* spread across all categories, potentially diluting their classification influence.

4.6 Analyzing mismatches

When running predictions on the test dataset, we capture the actual label, predicted label, and associated probabilities. We analyzed the probability distribution to determine if the predicted label was close in probability to the actual label. We observed that in cases of incorrect predictions, the predicted label was usually declared the winner with a high margin over other labels.
See prediction probabilities in the Appendix. In this example, we can see that only 2 of the 69 misclassified posts have a close margin of probability w.r.t the actual class prediction.

Examining the test and train datasets

Observation 1: Approximately 56% of the test dataset and 42% of the training dataset had class names that appeared in the corresponding posts, with an almost uniform ratio of appearance across all class types.
Observation 2: We found that only 3.6% of the test dataset and 3.5% of the training dataset contained synonyms for the class labels that appeared in the correct class.
Observation 3: Next, we examined the misclassified posts to see if the misclassified class name appeared against the misclassified label. About 17.5% of the test dataset showed the class name matching the wrongly predicted class label, while the presence of synonyms of the misclassified class name was negligible.

Discussion:
We believe that ML algorithms pick up on key tokens during fine-tuning, beyond what they already know from pre-trained data. We also believe that class names appeared frequently enough in these posts to be significant key tokens. In our test dataset, 56% of posts contained a class-name token matching the label, suggesting these tokens influenced correct classifications more than half the time. Even in mismatches, 18% of the class names aligned with misclassified labels (44 out of 245). We estimated that at least 65% of predictions were influenced by class-name tokens, and to verify this, we ran a masking test.

Other keywords like *suicide*, *death*, *anger*, and *frustration* likely influence predictions, although we hadn't analyzed their frequency or influence. Words like *anxiety* could be used in everyday speech rather than indicating a specific mental health condition, potentially skewing classifications.

4.7 Experiment 4 - Comparing social media posts with clinical definitions

Architecture and Design

Mental illness categories were formalized using definitions from the CDC, WHO, and Psychiatry.org. An NLP augmenter (BERT-base-uncased) generated 96 synthetic posts per category with neutral titles, which were combined with 96 original training posts for a balanced dataset. BERT was trained on this combined dataset and tested on the original dataset, excluding clinical definitions, achieving an accuracy of approximately 0.79.
For detailed results and confusion matrix, please see Reddit+CDC definitions.

Discussion

Incorporating clinical definitions into the training data reduced the accuracy compared to the baseline. This is likely because the clinical symptoms do not use the same colloquial language found in Reddit posts, even though they discuss similar symptoms. Despite the drop in accuracy, the results were not significantly affected. For future experiments, it is suggested that adding more data from mainstream sources like Wikipedia may help bridge the gap between clinical definitions and colloquial language in Reddit posts.

4.8 Experiment 5 - Masking class name tokens in posts

To assess the impact of class-name strings within the posts, we conducted experiments with different masking approaches. The baseline used a standard BERT classification with the original datasets, while Experiment A applied a mask only to the test set, and Experiment B masked both training and test datasets with dummy strings like mask1 or mask2.

Experiment	Train/Validation Set	Test Set	Overall accuracy
Baseline	Normal	Normal	0.84 (results)
A	Normal	Masked	0.74 (results)
B	Masked	Masked	0.78 (results)

Table 5: Experiment 5 results

Results and Discussion

As seen in Table 5, masking only the test set causes a 10% drop from the baseline, while masking both train and test sets results in a 6% drop. Experiment A shows that class name tokens assert an influence on the overall classification accuracy. However, experiment B shows that even without those class name tokens, the classifier is robust and can pick up cues from other contextual data.

5 Discussion of Evaluation Metrics

We have used overall F1 scores and Accuracy to quantify our classifier's performance. In our dataset, both micro and macro averages are likely to be the same because we have an equally

balanced set of classes. However, none of the aggregated metrics provide a granular view of per-class performance.
This is better served by looking at precision and recall for each of the classes. We see that the precision of the *depression* and *anxiety* labels does not perform as well as the others. This is in line with our observation done during our EDA.
The *bipolar* class has a relatively lower recall score. Looking at the confusion matrix, it seems that most bipolar cases are being misclassified as depression/ADHD/anxiety. The *ptsd* class has the next lowest recall score. Referring to the confusion matrix, it seems that most bipolar cases are being misclassified as depression/anxiety.

6. Conclusion and Future Work

We set out to create a generic classifier that can classify any Reddit posts and it can detect any of the indicators of mental illness. We were able to build such a classifier with a pre-trained RoBERTa model fine-tuned on our dataset of mental health-related subreddit posts. While we were able to replicate the work done by Murarta et al., we did not spend our efforts on improving it by tweaking NN architecture. But we instead focused on figuring out what could have possibly caused mismatches in classification.
We found that the presence of the class name string in the title/posts does influence the accuracy of the results. However, they are not as influential as we originally thought (based on Experiment B). We believe that other related words in the post can provide enough contextual information for the classifier to learn. Identifying such words would be certainly helpful in designing a generic classifier. For future research, we suggest that we start with a TF-IDF analysis of posts of different classes and then apply a masking approach to isolate words that matter.
We also experimented with feeding synthetic datasets based on CDC definitions of mental health symptoms into the training process. This unfortunately resulted in reduced accuracy on the test predictions. We believe that this is likely due to the limited set of synthetic training data. For future research, we suggest augmenting the dataset from other medical sources to build a larger corpus. We suggest training two classifiers in parallel - one that is based on clinical definition of symptoms and the other based on subreddit topics, both of which feed their logits to another higher level multi-class classifier which then learns through an intermediate dense layer.

References

1. Murarka, A., Radhakrishnan, B., & Ravichandran, S.(2021). Classification of mental illnesses on social media using RoBERTa. Proceedings of the 12thInternational Workshop on Health Text Mining and Information Analysis (https://arxiv.org/abs/2011.11226)
2. Vaswani, A., et al. (2017). Attention is all you need. 31st Conference on Neural Information Processing Systems. (https://arxiv.org/abs/1706.03762)
3. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. ArXiv, abs/1907, 11692. (https://arxiv.org/abs/1907.11692)
4. Nadeem A, Naveed M, Islam Satti M, Afzal H, Ahmad T, Kim KI. Depression Detection Based on Hybrid Deep Learning SSCL Framework Using Self-Attention Mechanism: An Application to Social Networking Data. Sensors (Basel). 2022 Dec 13;22(24):9775. doi: 10.3390/s22249775. PMID: 36560144; PMCID: PMC9782829. (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9782829/)

5. Jina Kim, Jieon Lee, Eunil Park, Jinyoung Han. A deep learning model for detecting mental illness from user content on social media. (https://pubmed.ncbi.nlm.nih.gov/32678250/)
6. Aldhyani THH, Alsubari SN, Alshebami AS, Alkahtani H, Ahmed ZAT. Detecting and Analyzing Suicidal Ideation on Social Media Using Deep Learning and Machine Learning Models. *International Journal of Environmental Research and Public Health*. 2022; 19(19):12635. (https://doi.org/10.3390/ijerph191912635)

Appendix

Same Hyperparameters as Murarka et al.

Murarka et al.	Models		posts			titles			posts+titles		
			P	R	F1	P	R	F1	P	R	F1
	RoBERTa	adhd	0.87	0.88	0.87	0.77	0.79	0.78	0.91	0.92	0.91
		anxiety	0.78	0.83	0.81	0.69	0.64	0.67	0.87	0.85	0.86
		bipolar	0.88	0.79	0.83	0.58	0.63	0.60	0.88	0.83	0.86
		depression	0.77	0.83	0.80	0.65	0.78	0.71	0.81	0.88	0.84
		ptsd	0.88	0.85	0.86	0.75	0.62	0.68	0.88	0.89	0.88
		none	0.99	0.95	0.97	0.94	0.88	0.91	1.00	0.98	0.99
Our Experiment	RoBERTa	adhd	0.88	0.80	0.84	0.83	0.75	0.79	0.81	0.93	0.87
		anxiety	0.84	0.70	0.76	0.59	0.71	0.64	0.82	0.88	0.85
		bipolar	0.67	0.82	0.74	0.60	0.63	0.62	0.83	0.73	0.77
		depression	0.70	0.82	0.86	0.64	0.75	0.69	0.86	0.81	0.84
		ptsd	0.87	0.83	0.85	0.75	0.56	0.64	0.89	0.87	0.88
		none	0.98	0.90	0.94	0.95	0.89	0.92	1.00	0.98	0.99

Table A1: RoBERTa Model - Comparison Test Results per Class-Same Hyperparameters as Murarka et al

Same Hyperparameters as Murarka et al.

Our Experiment	Models		posts			titles			posts+titles		
			P	R	F1	P	R	F1	P	R	F1
	DistilBERT	adhd	0.75	0.79	0.77	0.81	0.70	0.75	0.87	0.85	0.86
		anxiety	0.66	0.73	0.69	0.53	0.70	0.60	0.76	0.81	0.79
		bipolar	0.73	0.74	0.74	0.60	0.54	0.57	0.83	0.76	0.80
		depression	0.74	0.67	0.70	0.71	0.60	0.65	0.77	0.79	0.78
		ptsd	0.85	0.77	0.81	0.58	0.66	0.62	0.85	0.85	0.85
		none	0.94	0.96	0.95	0.93	0.89	0.91	0.97	0.98	0.98
	GPT2	adhd	0.82	0.80	0.81	0.71	0.72	0.71	0.89	0.85	0.87
		anxiety	0.73	0.72	0.73	0.59	0.69	0.64	0.78	0.82	0.80
		bipolar	0.80	0.75	0.78	0.61	0.54	0.57	0.77	0.79	0.78
		depression	0.67	0.83	0.75	0.65	0.73	0.69	0.75	0.85	0.80
		ptsd	0.91	0.76	0.83	0.67	0.60	0.64	0.92	0.80	0.86
		none	0.93	0.96	0.94	0.93	0.86	0.90	0.98	0.96	0.97
	XLM-RoBERTa	adhd	0.79	0.80	0.79	0.70	0.78	0.74	0.88	0.87	0.87
		anxiety	0.64	0.77	0.70	0.69	0.61	0.65	0.77	0.83	0.80
		bipolar	0.76	0.70	0.73	0.59	0.55	0.57	0.83	0.76	0.79
		depression	0.77	0.72	0.74	0.63	0.71	0.67	0.78	0.85	0.82
		ptsd	0.88	0.79	0.83	0.64	0.66	0.65	0.87	0.81	0.84
		none	0.93	0.96	0.94	0.96	0.87	0.91	0.97	0.98	0.97

Table A2: DistilBERT, GPT2, XLM-RoBERTa (Same Hyperparameters as Murarka et al.)

EDA on the training, validation and test datasets

Balanced Training/Validation Datasets

EDA showed that the classes are fairly balanced on both the training and validation datasets.

```
Training set
Label adhd in training set  : 2465/13727 (17.96%)
Label anxiety in training set  : 2422/13727 (17.64%)
Label bipolar in training set  : 2407/13727 (17.53%)
Label depression in training set  : 2450/13727 (17.85%)
Label ptsd in training set  : 2001/13727 (14.58%)
Label none in training set  : 1982/13727 (14.44%)

Validation set
Label adhd in validation set: 248/1488 (16.67%)
Label anxiety in validation set: 248/1488 (16.67%)
Label bipolar in validation set: 248/1488 (16.67%)
Label depression in validation set: 248/1488 (16.67%)
Label ptsd in validation set: 248/1488 (16.67%)
Label none in validation set: 248/1488 (16.67%)

Test set
Label adhd in test set: 248/1488 (16.67%)
Label anxiety in test set: 248/1488 (16.67%)
Label bipolar in test set: 248/1488 (16.67%)
Label depression in test set: 248/1488 (16.67%)
Label ptsd in test set: 248/1488 (16.67%)
Label none in test set: 248/1488 (16.67%)
```

Presence of class names in title+posts

We examined the presence of the class names in the body of the posts. Although the posts had been labeled based on subreddit topics, several posts carry the class words (e.g. *depression, adhd* etc) in the body of the post as well. A sample of these posts can be found here:

Below is a summarized stats on the training posts.

Prevalence of class names in title+posts (training)

Labeled class:	adhd	depression	anxiety	ptsd	none	bipolar	absent in post
Token adhd	51%	0%	0%	0%	0%	2%	47%
Token anxiety	6%	6%	45%	9%	0%	8%	27%
Token bipolar	1%	0%	1%	1%	0%	34%	63%
Token depression	5%	23%	6%	5%	0%	13%	48%
Token ptsd	0%	1%	1%	49%	0%	1%	48%
Token none	0%	0%	0%	0%	0%	0%	100%

Interpretation: The percentage chart shows the ratio of posts with the label name to the total posts for that label. For example, *adhd* appears in 1285 of 2465 ADHD-labeled posts, or 51%. The absolute numbers behind those percentage charts are shown below

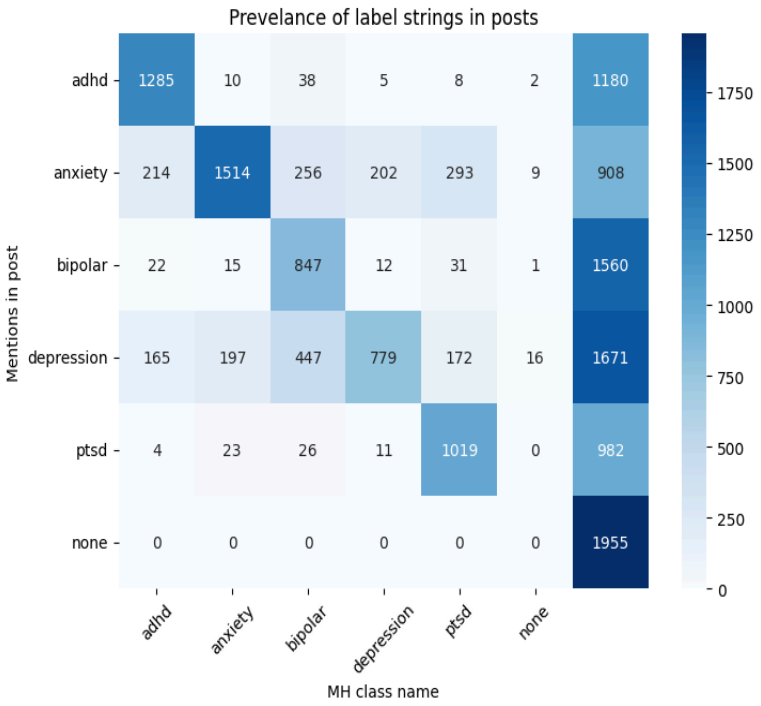
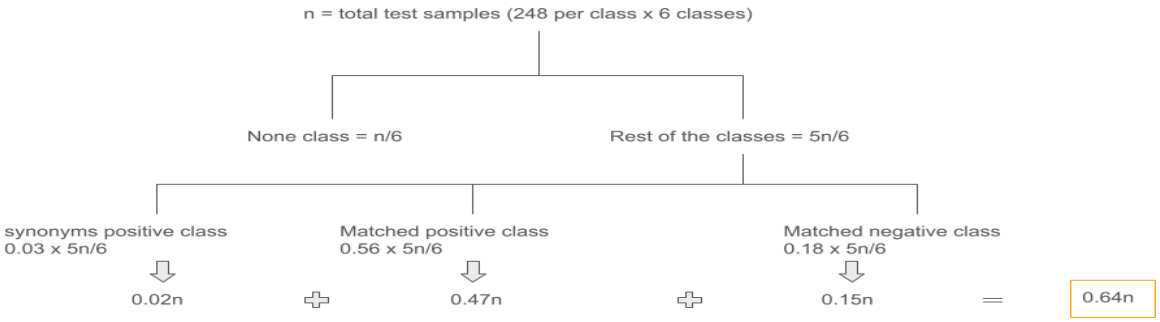


Figure A1: Prevalence of class names in posts

Observation: Class name words like *adhd*, *bipolar*, and *ptsd* are concentrated in their respective categories, while *depression* and *anxiety* spread across all categories, potentially diluting their classification influence.

Test results explainable by matching classnames to class labels



Probability predictions in misclassified dataset

```
#@title Inspect the result dataframe
errs = res_df[(res_df['actual'] == 'bipolar') & (res_df['predicted'] != 'bipolar') ]
errs.sort_values(by='prob_bipolar', ascending=False)
```

	actual	predicted	test_id	text	prob_adhd	prob_anxiety	prob_bipolar	prob_depression	prob_ptsd	prob_none
237	bipolar	depression	237	it's my 31st birthday i didn't care to do anyt...	0.078031	0.018507	0.417828	0.455235	0.017411	0.012989
228	bipolar	ptsd	228	taking ownership do any of you have or had tro...	0.014099	0.072408	0.400505	0.060723	0.412998	0.039267
1332	bipolar	depression	1332	for the first time since i was 9 i can say i a...	0.010555	0.009596	0.368446	0.601927	0.005757	0.003720
983	bipolar	anxiety	983	at what point is it psychosis? over the past y...	0.035321	0.433761	0.360509	0.047731	0.086805	0.035872
337	bipolar	depression	337	louisiana flood victim, and i'm episodic i don...	0.028041	0.014946	0.356986	0.574532	0.020024	0.005471
...	...	...	...	...	...	...	...	...	...	...
368	bipolar	anxiety	368	my new cat i got a cat almost a month ago. i g...	0.006003	0.977705	0.004388	0.005301	0.005510	0.001093
1294	bipolar	ptsd	1294	do you guys get nightmares? i had the worst on...	0.000049	0.000423	0.000775	0.002167	0.996254	0.000333
65	bipolar	ptsd	65	forfeited all my attendance points today. thir...	0.000102	0.000119	0.000455	0.000551	0.998102	0.000670
187	bipolar	ptsd	187	therapy during episodes. what do you do and is...	0.000011	0.000013	0.000136	0.000193	0.999595	0.000052
741	bipolar	ptsd	741	i cried today, but out of appreciation. so, as...	0.000013	0.000014	0.000128	0.000160	0.999602	0.000084

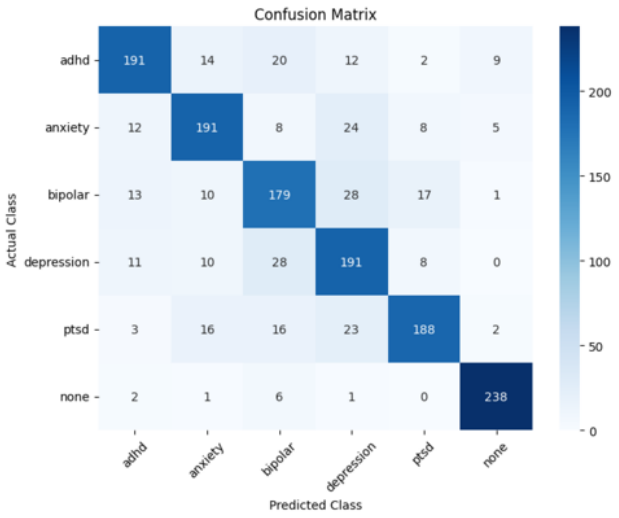
69 rows × 10 columns

Confusion matrix with Reddit post and clinical definitions training set

In this experiment, we combine clinical definitions texts as ‘posts’ with a neutral title like ‘symptoms defined by CDC’. This is mixed with actual posts and titles from labeled subreddit posts. The following are the results:

Accuracy: 0.7916666666666666

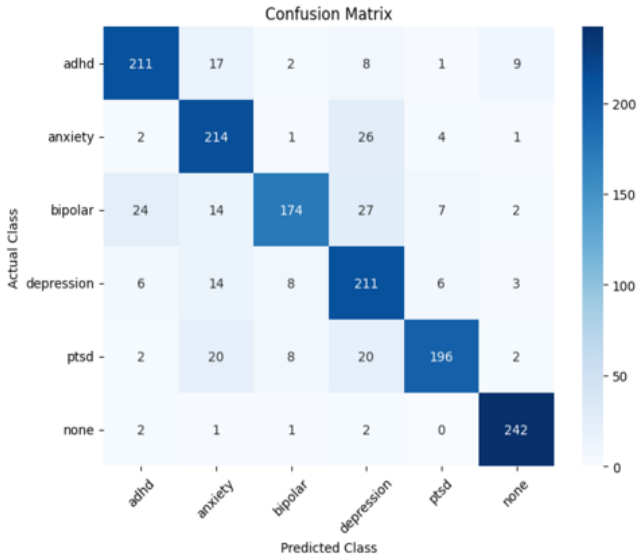
	precision	recall	f1-score	support
adhd	0.82	0.77	0.80	248
anxiety	0.79	0.77	0.78	248
bipolar	0.70	0.72	0.71	248
depression	0.68	0.77	0.72	248
ptsd	0.84	0.76	0.80	248
none	0.93	0.96	0.95	248
accuracy			0.79	1488
macro avg	0.80	0.79	0.79	1488
weighted avg	0.80	0.79	0.79	1488



Baseline Metrics before masking

Accuracy: 0.8387096774193549

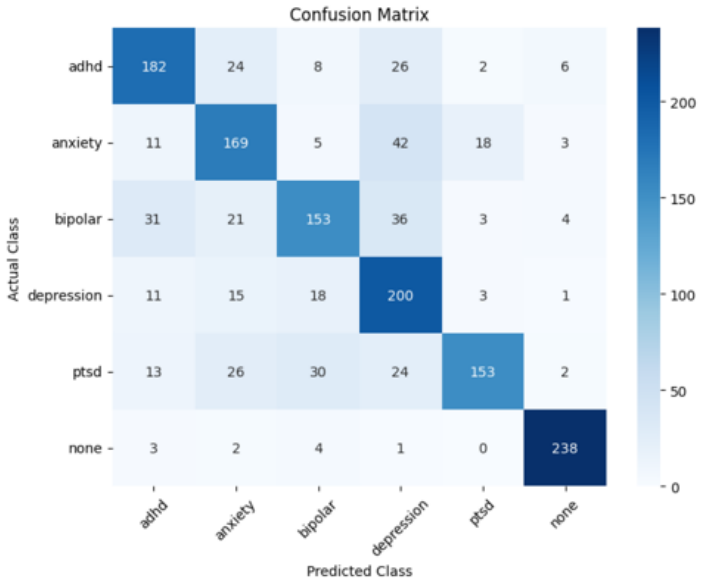
	precision	recall	f1-score	support
adhd	0.85	0.85	0.85	248
anxiety	0.76	0.86	0.81	248
bipolar	0.90	0.70	0.79	248
depression	0.72	0.85	0.78	248
ptsd	0.92	0.79	0.85	248
none	0.93	0.98	0.95	248
accuracy			0.84	1488
macro avg	0.85	0.84	0.84	1488
weighted avg	0.85	0.84	0.84	1488



Experiment A (masking test dataset only)

Accuracy: 0.7358870967741935

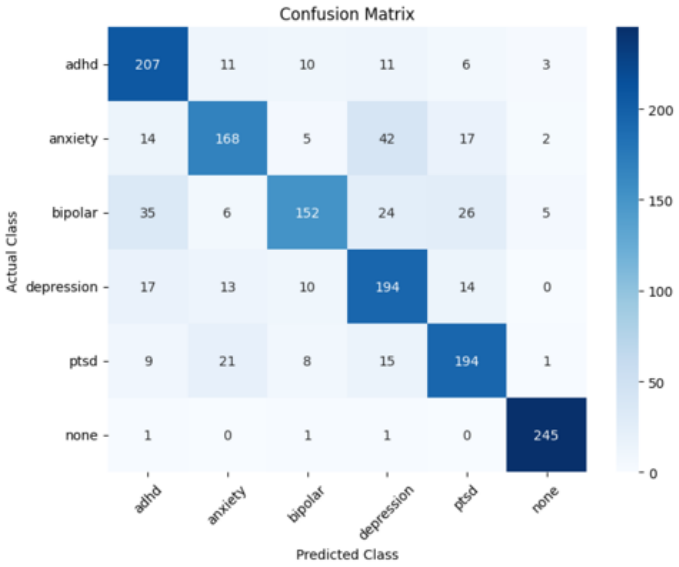
	precision	recall	f1-score	support
adhd	0.73	0.73	0.73	248
anxiety	0.66	0.68	0.67	248
bipolar	0.70	0.62	0.66	248
depression	0.61	0.81	0.69	248
ptsd	0.85	0.62	0.72	248
none	0.94	0.96	0.95	248
accuracy			0.74	1488
macro avg	0.75	0.74	0.74	1488
weighted avg	0.75	0.74	0.74	1488



Experiment B (masking train & test datasets)

Accuracy: 0.7795698924731183

	precision	recall	f1-score	support
adhd	0.73	0.83	0.78	248
anxiety	0.77	0.68	0.72	248
bipolar	0.82	0.61	0.70	248
depression	0.68	0.78	0.73	248
ptsd	0.75	0.78	0.77	248
none	0.96	0.99	0.97	248
accuracy			0.78	1488
macro avg	0.78	0.78	0.78	1488
weighted avg	0.78	0.78	0.78	1488



Classname matches in test dataset

```
Matches for string adhd in true class adhd : 215
Matches for string depression in true class depression : 204
Matches for string ptsd in true class ptsd : 211
Matches for string anxiety in true class anxiety : 196
Matches for string bipolar in true class bipolar : 179

Total Matches                              : 1005

Occurance of string matching (actual) classname in posts  : 565
Percentage of string matching (actual) classname in posts : 56.21
```

Classname matches in train dataset

```
Matches for string adhd in true class adhd : 2465
Matches for string depression in true class depression : 2450
Matches for string ptsd in true class ptsd : 2001
Matches for string anxiety in true class anxiety : 2422
Matches for string bipolar in true class bipolar : 2407

Total Matches                              : 11745

Occurance of string matching (actual) classname in posts  : 4914
Percentage of string matching (actual) classname in posts : 41.84
```

Misclassified class name in predicted class

```
Misclassified classname string found in predicted class; True Class: adhd                : 33
Misclassified classname string found in predicted class; True Class: depression        : 44
Misclassified classname string found in predicted class; True Class: ptsd                :
37
Misclassified classname string found in predicted class; True Class: anxiety             : 52
Misclassified classname string found in predicted class; True Class: bipolar             : 69
Total mismatches                           : 235
```

```
Occurance of mismatched (predicted) classname  : 41
Percentage of mismatched (predicted) classname : 17.4468085106383
```

Granular metrics on classifications based on different language models (Updated Hyperparameters)

Model	Class	posts			titles			posts+titles		
		P	R	F1	P	R	F1	P	R	F1
BERT	adhd	0.89	0.75	0.81	0.71	0.77	0.74	0.83	0.88	0.85
	anxiety	0.66	0.83	0.73	0.62	0.68	0.65	0.81	0.82	0.81
	bipolar	0.79	0.77	0.78	0.63	0.51	0.56	0.86	0.79	0.83
	depression	0.73	0.76	0.75	0.63	0.72	0.68	0.84	0.79	0.81
	ptsd	0.85	0.78	0.82	0.70	0.59	0.64	0.84	0.88	0.86
	none	0.96	0.92	0.94	0.89	0.91	0.90	0.96	0.97	0.96
RoBERTa	adhd	0.80	0.79	0.79	0.69	0.69	0.69	0.90	0.83	0.86
	anxiety	0.70	0.78	0.74	0.60	0.68	0.64	0.82	0.83	0.82
	bipolar	0.80	0.72	0.76	0.79	0.40	0.53	0.94	0.68	0.79
	depression	0.66	0.79	0.72	0.55	0.78	0.65	0.71	0.89	0.79
	ptsd	0.92	0.75	0.83	0.67	0.65	0.66	0.89	0.83	0.86
	none	0.95	0.92	0.94	0.90	0.88	0.89	0.87	0.99	0.92
DistilBERT	adhd	0.81	0.74	0.77	0.82	0.59	0.69	0.80	0.85	0.82
	anxiety	0.61	0.77	0.68	0.68	0.62	0.65	0.80	0.79	0.79
	bipolar	0.87	0.57	0.69	0.47	0.70	0.56	0.76	0.79	0.77
	depression	0.63	0.76	0.69	0.62	0.73	0.67	0.73	0.83	0.78
	ptsd	0.80	0.78	0.79	0.79	0.54	0.65	0.96	0.71	0.82
	none	0.95	0.95	0.95	0.90	0.92	0.91	0.93	0.98	0.96
GPT2	adhd	0.72	0.84	0.77	0.62	0.69	0.66	0.92	0.79	0.85
	anxiety	0.81	0.69	0.75	0.64	0.62	0.63	0.86	0.89	0.82
	bipolar	0.72	0.73	0.73	0.62	0.42	0.50	0.81	0.73	0.81
	depression	0.66	0.70	0.68	0.56	0.73	0.63	0.79	0.79	0.78
	ptsd	0.87	0.79	0.83	0.65	0.63	0.64	0.91	0.82	0.84
	none	0.93	0.94	0.93	0.89	0.87	0.88	0.95	0.96	0.96
XLM-RoBERTa	adhd	0.71	0.77	0.74	0.68	0.68	0.68	0.83	0.85	0.84
	anxiety	0.63	0.73	0.68	0.62	0.60	0.61	0.80	0.82	0.81
	bipolar	0.67	0.67	0.67	0.57	0.49	0.53	0.79	0.74	0.77
	depression	0.65	0.68	0.66	0.58	0.68	0.63	0.75	0.83	0.79
	ptsd	0.83	0.71	0.76	0.59	0.60	0.59	0.83	0.82	0.82
	none	0.96	0.84	0.90	0.88	0.87	0.87	0.99	0.90	0.94

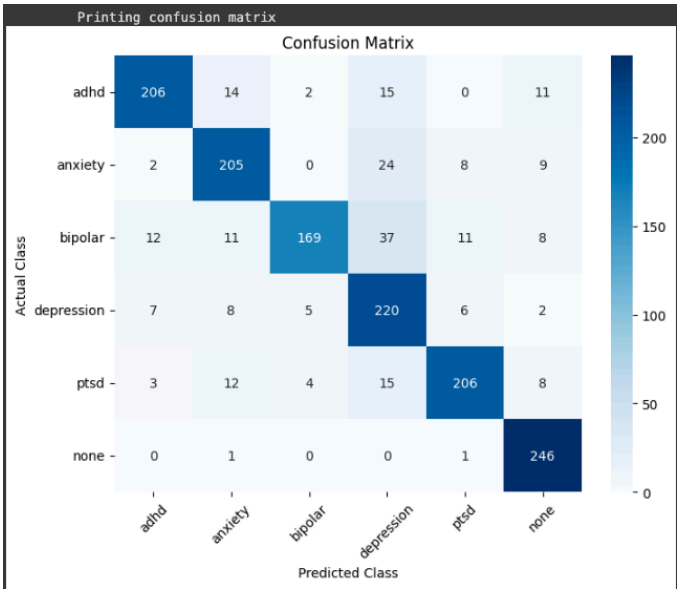
Table A3: Model results executed with updated hyperparameters

Recreating Baseline results

```
Accuracy: 0.8413978494623656
              precision    recall  f1-score   support

        adhd       0.90      0.83      0.86       248
     anxiety       0.82      0.83      0.82       248
     bipolar       0.94      0.68      0.79       248
  depression       0.71      0.89      0.79       248
        ptsd       0.89      0.83      0.86       248
        none       0.87      0.99      0.92       248

    accuracy                           0.84      1488
   macro avg       0.85      0.84      0.84      1488
weighted avg       0.85      0.84      0.84      1488
```



Sample Posts containing class-names embedded in the post

Label	Sample Post
Bipolar	Like tracy is so bipolar, i really don't like her. she's so mean ." no she's just a bitch it's completely different. or im really struggling today. dont want to tale my meds. dont want to be bipolar. want to be able to handle my emotions ra basically i just feel like i cycle through these very distinct but very different personalities on a day to day basis. i hi all. i have recently been diagnosed with bipolar affective disorder.…
Anxiety	working gives me so much anxiety but i can't rely on others financially anymore. i wish i didn't have to work but then i i'm having some pretty bad anxiety right now about my brother dying :/ it's usually always about my parents but it doesn i finally decided to go to my doctor and tell them about my anxiety - currently in the waiting room having anxiety about do you ever feel like what makes your decisions is your anxiety? ….
depression	users of r/depression i ask you to upvote peoples posts. it might make all the difference and make them feel slightly be there's nothing really bad every happened to me in the past or the now but my future is what is one of the major reasons it's been a disgusting mess for over a year. i'm tired of it. it really does add to my overall stress, depression and an i have had this feeling for some time now. my depression doesn't seem to go away. problems are not receding. i just wann they don't seem to understand that depression isn't something that affects me at random. i deal with it daily. they don' "all it takes is a beautiful smile to hide how broken you are." depression to me is looking for the best place to hang…
ptsd	pretty much the question. i'm having a really hard time with what to do. i love my job, but can't get away from bullying does ptsd make anyone else crazy dizzy? is this a common thing? or is it because the incidient itself made me really diz i find that any time i have a physical / medical issue it's almost always first blamed on ptsd, anxiety, depression, or i think working out is suppose to help, but i get panic attacks i can't come back from after working out. none of my pts i know ptsd and my past experi…
adhd	a facebook friend posted a quote "adhd isn't something in your child that needs to be fixed. it's a superpower they need honestly an adhd person arguing a point is like when your exploring minecraft . it's all fine and good until the chunks literally the above. paid $15 at the pump and then drove off, forgetting to fill the tank. …

Sample Titles containing class-names embedded in the titles

Label	Sample Post
Bipolar	have any of you single bipolarbears found a way to control your finances? does anyone else hate when people use the word bipolar like it's some horrible thing?
Anxiety	anxiety is not a choice. people who laugh, yell, and mock the people with anxiety and panic disorders are sad people. anxiety problems hiding when a stranger comes to the door holy sheeeet i just emailed all my soon-to-be professors about my situation and i'm having anxiety attacks over it but i
depression	i'm an artist. i dont like talking about my depression much so i drew it. do you think depression has reduced your mental abilities and affected your decision making skills? i shouldn't have to lie about my depression to get a sick day
ptsd	neglect ptsd do you actually use the word "ptsd" when explaining your situation? marijuana changed my ptsd
adhd	a glimpse into an adhd brain i actually don’t have adhd and i’m not going back to law school. thank you so much for everything r/adhd i'm just another student with adhd who is completely floored that i got into college.