Categorizing Mental Health Through Reddit Posts During the COVID-19 Pandemic

Stanford CS224N Custom Project
TA mentor: Elaine Sui

Elizabeth Fitzgerald
Department of Computer Science
Stanford University
elizfitz@stanford.edu

Natasha Saki Ong
Department of Computer Science
Stanford University
natashao@stanford.edu

Abstract

With our project, we seek to categorize Reddit posts by the mental health disorders they indicate. We rely upon the use of emotion labels to accurately categorize mental health disorders from Reddit posts. We experimented with 4 different models that varied in how they used the text posts and their associated emotions. Overall, we found that the model which performed best was Model 4. This model was trained in 2 stages, the first of which classified emotion labels. We then took this trained model, altered the classification layer, and retrained this final layer to predict mental health conditions. The model performed fairly well, achieving an F1-score of 0.7 for the majority of conditions. Having found a model we were satisfied with, we completed trend analysis on new Reddit data. This revealed that, as we suspected, the manifestation of certain mental health disorders, especially Depression and Anxiety, changed over the course of the pandemic.

1 Introduction

Identifying and tracking mental health trends online has never been more important than during the COVID-19 pandemic. While there is a significant amount of literature surrounding Reddit and mental health, there is very little that explores how it has changed over the last two years. This brings us to our primary goal. With this project, we seek to determine how the online presence of common mental health disorders was impacted during different stages of the pandemic. We find this problem interesting for 2 main reasons: (1) there is so much public data from social media and it would be useful to leverage this data for this application, (2) there is little data on mental health shifts over the course of the pandemic (still ongoing at the time of writing this paper).

Our project relies on being able to classify mental health conditions, and we experimented with different components of the training pipeline. First, we determined which of two datasets to use: an existing dataset composed of Reddit posts, and a custom dataset of Reddit comments. Ultimately, we decided to work with the post data due to the better performance in being able to distinguish between different mental health disorders. Comments seemed to be more similar across subreddits, thus making it difficult to capture the nuance of each condition compared to posts. We labeled each post both for the mental health condition it indicated, as well as the emotions it presented. Next, we used this data in four experiments with model architecture. These four experiments varied in input variables, output variables, and structure. All, however, relied upon fine-tuning a pre-trained BERT model. Once all four models were evaluated, we used the most successful to predict and analyze the mental health conditions of "neutral" posts distributed evenly across the years 2020 to 2022.

This is a particularly challenging problem. Collecting the right data and information is already hard, as it is difficult to even claim that mental health disorders can be identified from comments alone. Beyond that, predicting the emotions present in text (a limited form of emotional analysis) is also an entire sub-field of NLP. Finally, potentially combining the two to gain insights on internet mental

health trends raises questions of efficacy. While we feel we have made a good faith effort to tackle this challenge head-on, we feel we must acknowledge how difficult it was to do so in only 10 weeks.

2 Related Work

There is a variety of prior work related to our project. To start, even before the COVID-19 pandemic, researchers have worked on using social media to study mental health trends on individual and population levels. In particular, researchers have attempted to identify quantifiable signals in Tweets for bipolar disorder, major depressive disorder, PTSD and seasonal affective disorder [1]. They construct a dataset by identifying self-expressions of mental illness diagnoses and build classifiers to separate diagnosed users from control users by using 2 types of language models. One thing to note is that their dataset relies on self-identification of a diagnosis of a condition, which may be rarer to find than leveraging general social media content. Our project overcomes these limitations by leveraging a larger amount of the publicly available data (ie we do not filter content for specific self-identifying terms) and relying upon existing state-of-the-art models (ie BERT) to further improve performance. However, our data is, in a sense, less accurate because we have no explicit confirmation that a post is representative of a user with any given mental health disorder.

However, since the onset of the pandemic, there have been a number of studies that have conducted mental health analyses using publicly available data. For instance, some researchers have analyzed mental and physical disorders associated with COVID-19 in online health forums during the early pandemic – January to May 2020 [2]. The authors of this paper found that mental health symptom keywords were more frequently mentioned by authors of COVID-19-related posts as compared to physical health symptom keywords. The authors correlate the peak in mental and physical health keyword mentions to when the World Health Organization declared the COVID-19 pandemic. This paper, though insightful, only covers the first 4 months of the pandemic and does not use deep learning techniques. It also uses data taken from online health forums, which is a narrow selection of text data. In contrast, our work investigates trends in neutral subreddits and looks at two years of the pandemic.

Some researchers have already attempted to classify mental health disorders from Reddit posts and investigate features that characterize each disorder as well as whether particular subreddits were becoming more similar to each other over the pandemic [3]. The authors were thorough in their approach, as they experimented with used a variety of supervised learning techniques, as well as keyword analysis to look for trends over time. We take inspiration from this paper, particularly in our decision to use a dataset with Reddit posts (as opposed to comments, a decision we will discuss in Section 4.1) and using posts from "neutral" subreddits to do our analysis of trends. However, like the other papers, this paper also fails to use deep learning models. Moreover, the authors used the model with the lowest complexity in determining feature importance, which may be problematic. Our project attempts to use deep learning models, specifically fine-tuned BERT models, to do classification on the text data itself, as we do not have the background in psychology necessary to make subjective choices on text features that may indicate a mental health disorder. Moreover, our project investigates using other data (i.e. emotion labels) as further inputs to our model to further improve model performance.

3 Approach

While the specifics of each model varied, all four of our experiments are built on a pre-trained BERT model (bert-base-uncased) and fine-tuned by adding two linear layers (dimensions: (768, 256) and (256, 6)) separated by a ReLU layer. Because initial performance for Models 2 and 4 was promising, we sought to improve their results further by adding a dropout layer after the ReLU layer. Further, AdamW optimizer is used. A diagram of the general model architecture can be seen in Figure 1 below. For each of our four models, we did hyperparameter tuning. This involved an exhaustive search of combinations of learning rates, batch sizes, and epochs. The results of the hyperparameter search can be seen in Section 4.3.

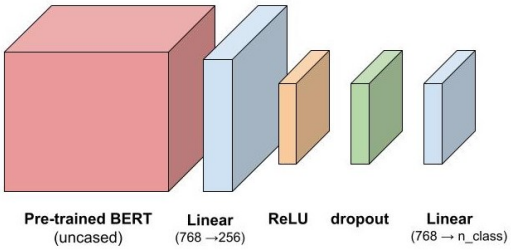

Figure 1: General Model Architecture

Ultimately, the goal of our project was to create a model with which we could confidently do mental health trend analysis on a selection of subreddits. In order to do this, we hypothesized that our best model would need to rely on emotions. Specifically, our approach relied upon different methods of encoding emotions into our models. Our basis for this decision was the possibility that emotions correlate differently with certain mental health disorders. We base this claim on the criteria for mental health diagnoses in the DSM-5 ([4]). Generally speaking, we were relying on the assumption that Reddit posts will not only communicate the emotions a person is really feeling, but also that those emotions are indicative of a mental health disorder.

In light of this, our baseline model did not rely upon emotions, and the only input was the text data. The experiment details are listed in Section 4.3, but Experiment 1 represents how well a fine-tuned, pre-trained BERT model can do on the task without the consideration of emotions.

4 Experiments

4.1 Data

Over the course of this project, we have grappled with the problem of knowing what data to use. **We will describe both datasets we considered, although we ultimately chose to use the post data from the Reddit Mental Health Dataset [5] for model training.** Our analysis data was scraped separately, as will be described later.

Our first attempt to get data was to create our own dataset of Reddit comments by scraping the Reddit API with using the Pushshift API wrapper, PMAW ([6]). Our manual dataset is comprised of 180,000 randomly sampled comments which fall into six classes (according to the subreddits we pulled them from). There are 30,000 comments each from the following subreddits: r/depression, r/Anxiety, r/adhd, r/BipolarReddit, and r/addiction. Finally, there is an additional 30,000 comments representing the "No Mental Health Disorder" category, which is made up of 10,000 comments from each of r/AskReddit, r/Showerthoughts, and r/jokes. All comments were scraped randomly from between the dates 01/01/2020 and 01/01/2022.

However, making this custom dataset proved challenging and time consuming. As a backup option, we searched for public datasets, and found the Reddit Mental Health Dataset[5]. The dataset is significantly imbalanced, with the addiction and Bipolar disorder classes having roughly 5000-6000 posts while the Anxiety class had roughly 60,000 posts. To combat this imbalance, class weights were used when training each model. Each of the 5 condition classes (Anxiety, Depression, Bipolar Disorder, Addiction, ADHD) are represented as with our custom dataset, with the non-mental-health-condition dataset comprised of posts from r/jokes and r/fitness. According to the authors, these posts were scraped from the beginning of 2018 to the middle of 2020.

Regardless of dataset, the class labels are pseudolabels, ie the names of mental health disorders associated with each subreddit, or "none." To obtain our emotion label, we used the NRC Word-Emotion Association Lexicon [7], which is a list of English words and their associations with 8 basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust). In order to determine which emotions were present in our data, we cross-referenced words used in the text with words listed in the lexicon. If the word was found in the lexicon, then we marked the emotion associated with that word as being present. Both condition and emotion labels are one-hot encoded. Regardless of the data, we used a train-test-dev split of 80-10-10.

Finally, for our trend analysis dataset, we relied on the same methodology used to generate our custom dataset to scrape post data from three neutral subreddits: r/confessions, r/legaladvice, and r/personalfinance. This data is split into eight, three-month chunks of time starting on 01/01/2020 and going until 01/01/2022. Each time period contains 1,000 posts from each of the subreddits, for a total of 24,000 posts for analysis.

4.2 Evaluation method

To evaluate the performance of each model prior to doing the trend analysis, we compare F1-scores and confusion matrices. F1-score is defined as the harmonic mean of Precision and Recall. In calculating the F1-scores for each class, we calculated the precision and recall for each class. The general precision, recall, and F1-scores have formulas given by:

$$\text{Precision} = \frac{\text{TP}}{\text{TP + FP}} \qquad \text{Recall} = \frac{\text{TP}}{\text{TP + FN}}$$

$$F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision + Recall}} = \frac{\text{TP}}{\text{TP} + \frac{1}{2}(\text{FP + FN})}$$

Here, TP, FP, TN, FN represent the true positive (TP), false positive (FP), true negative (TN), and false negative (FN) for the relevant classes. Although accuracy is a simple and widely utilized way to measure model performance with regards to the model's ability to make correct predictions, we recognize that accuracy is not necessarily a trustful way to evaluate our model. We care about precision and recall as both costs of false positives and false negatives are high. In particular, to be able to reliably make conclusions from our trend analysis, we need to make sure our model is sensitive and precise. The more false positives and false negatives there are, the more false trends we might be detecting. As such, the F1-score, which combines both precision and recall, is most appropriate for our task.

In particular, we compare the F1-scores for individual classes and investigate the types of classification errors the models make. In terms of determining which model is better, consider this hypothetical. Model A returns an F1-score of around 0.60 for each mental health disorder. Model B returns an F1-score of 0.95 for three of the disorders, but only 0.10 for the others. Perhaps Model B is simply categorizing everything as one of those three disorders. In this case, we would prefer Model A.

4.3 Experimental details

We experiment with 4 different ways of utilizing the post data, condition labels, and emotion labels. A visual representation of our four experiments can be seen in Figure 2.

- **Experiment 1 (baseline)**: We fine-tune the BERT model from Section 3 on Reddit posts to output the mental health condition label. Directly training the model on text to output labels is a pretty common procedure and is thus used as our baseline. After hyperparameter tuning, this model uses Pytorch's Cross Entropy Loss with $lr = 5e-5, \epsilon = 1e-8, \beta_1 = 0.99, \beta_2 = 0.999$, batch size of 32 and 2 epochs. No dropout is applied.
- **Experiment 2**: We fine-tune the BERT model from Section 3 on both Reddit posts and the emotions labels to predict the mental health condition label. To do so, we concatenate the emotion labels represented as text (ex. "fear anger" instead of one-hot vector) and the original Reddit post. Since the BERT model can consume 512 tokens maximum and truncates anything beyond that, we concatenate the emotion labels to the beginning of the Reddit post. This ensures that even if a given Reddit post is long (ie gets truncated), the model is fed some data about the emotions present. After hyperparameter tuning, we used a Pytorch's Cross Entropy Loss with $lr = 1e-3, \epsilon = 1e-8, \beta_1 = 0.99, \beta_2 = 0.999$, batch size of 32 and 3 epochs. Since initial results were promising, we also use dropout of 0.2
- **Experiment 3**: We fine-tune the BERT model from Section 3 on Reddit posts to predict both the mental health condition and the emotions present in the text. Since the output is multi label (conditon and emotion), we use Pytorch's Binary Cross Entropy loss instead of Cross Entropy Loss, which uses a sigmoid instead of a softmax. After hyperparameter tuning, we use Pytorch's Binary Cross Entropy Loss with $lr = 5e-4, \epsilon = 1e-8, \beta_1 = 0.99, \beta_2 = 0.999$, batch size of 64 and 1 epochs. No dropout is applied.

- **Experiment 4**: This model is our most sophisticated model and consists of 2 stages. We start by fine-tuning the BERT model from Section 3 on Reddit posts to output emotion labels. For this, we used a Binary Cross Entropy Loss function since we are dealing with multi-label classification. After hyperparameter-tuning for this stage, we use $lr = 5e-4, \epsilon = 1e-8, \beta_1 = 0.99, \beta_2 = 0.999$, batch size of 64 and 2 epochs. We saved this model, dropped the last layer, and added a new, randomly initialized layer with output dimension of 6 for the 6 conditions (5 mental health conditions and "none"). We switch to Pytorch's Cross Entropy Loss for this step. After hyperparameter-tuning, our model used $lr = 1e-3, \epsilon = 1e-8, \beta_1 = 0.99, \beta_2 = 0.999$, batch size of 32 and 3 epochs. Initial results were promising and there was slight overfitting, so a dropout of 0.2 was used.

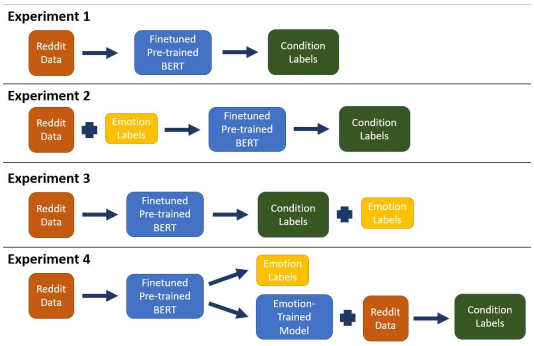

Figure 2: Experiment Designs Overview

4.4 Results

4.4.1 Quantitative Results

Experiment 1 performed surprisingly well. Precision, recall, and F1-scores were calculated for each class and can be seen in Table 1. Likewise, the generated confusion matrix can be seen in Figure 5. Training directly on posts thus seems to be quite effective for those with limited resources (data, time, etc).

Experiment 2 outperformed the baseline and precision, recall, and F1-scores for each class can be seen in Table 2. Likewise, the generated confusion matrix can be found in Figure 6. Compared to the baseline, we see that almost every metric improved. The one exception is the recall for Depression, which decreased by 0.02. Classifications for Bipolar Disorder saw the most improvement, although based upon the confusion matrix, this may be because the model simply predicted the condition more frequently.

Experiment 3 tested whether training a model to predict both the condition and the emotions would improve performance only with regards to classifying mental health disorders. Again, precision, recall, and F1-scores can be seen in Table 3. Likewise, the generated confusion matrix can be seen in Figure 7. While the precision scores improved from the baseline, sometimes greatly, the recall scores dropped drastically. The reasoning for this becomes clear when looking at the confusion matrix. Generally speaking, the model made far fewer predictions other than Depression. Thus, we can say that while we can feel more confident in the model's predictions for Anxiety, Bipolar, and Addiction, we know it is misclassifying many of the posts for those conditions as Depression. This is the sort of imbalance we would like to avoid for our best model.

Finally, Experiment 4 tested whether training a model to predict both the condition and the emotions would improve performance only with regards to classifying mental health disorders. Again, precision, recall, and F1-scores for each class can be seen in Table 4. Likewise, the generated confusion matrix can be seen in Figure 8. Relative to both Experiments 1 and 2, we see that almost every metric improved. The improvements are not extreme, but they are evenly showcased across all conditions, which is promising. Like with Experiment 2, the classifications for Bipolar Disorder showed the most improvement. However, the confusion matrix for Model 4 seems to imply that this was not simply because it predicted Bipolar Disorder more frequently. Rather, it simply made more confident

guesses, although Model 4's recall score for Bipolar Disorder was marginally worse than that of Model 2.

Condition	Precision	Recall	F1
Depression	0.74	0.73	0.73
Anxiety	0.71	0.67	0.69
Bipolar	0.23	0.29	0.26
Addiction	0.41	0.75	0.53
ADHD	0.70	0.66	0.68
None	0.94	0.92	0.93

Table 1: Exp. 1 - Precision/recall/F1-score

Condition	Precision	Recall	F1
Depression	0.76	0.71	0.73
Anxiety	0.72	0.70	0.71
Bipolar	0.26	0.41	0.32
Addiction	0.47	0.78	0.59
ADHD	0.74	0.67	0.70
None	0.95	0.96	0.96

Table 2: Exp. 2 - Precision/recall/F1

Condition	Precision	Recall	F1
Depression	0.77	069	0.73
Anxiety	0.71	0.67	0.69
Bipolar	0.58	0.03	0.05
Addiction	0.85	0.45	0.59
ADHD	0.80	0.55	0.65
None	0.96	0.93	0.94

Table 3: Exp. 3 - Precision/recall/F1-score

Condition	Precision	Recall	F1-score
Depression	0.78	0.73	0.76
Anxiety	0.73	0.74	0.73
Bipolar	0.29	0.39	0.33
Addiction	0.47	0.80	0.59
ADHD	0.78	0.70	0.73
None	0.96	0.96	0.96

Table 4: Exp. 4 - Precision/recall/F1-score

5 Analysis

5.1 Qualitative Results Analysis

Overall, these quantitative results led us to use **Model 4** for our trend analysis. We felt it showed the most improvement, particularly over Model 2. Unfortunately, Experiment 3 demonstrated that the predicting the emotions did not benefit the condition predictions, although it served as a useful point of comparison for Experiments 2 and 4.

To better understand the systems, we looked at some examples of misclassified texts/samples from Model 2 and 4 (our two most promising models). Table 5 has few wrongly classified examples from Model 2. In the first example, we see that the model's output isn't entirely wrong from what it is provided. More specifically, Bipolar condition is characterized by episodes of mania and depression [8], but the model can only see the depression aspect of this condition in this post. As such, given the context and data the model has access to, it makes sense why the model might predict depression instead of Bipolar disorder. In the second example, the term "addiction" appears in the post but the model doesn't predict addiction. This means that our model doesn't predict addiction just because of the presence of the term "addiction" itself. The model might have predicted a label other than addiction because the majority of texts in the dataset for addiction relate to drinking or drug addiction instead of food-related addictions. For the third example, we see that the model outputted ADHD instead of Addiction. This text is hard to classify without knowing who Keith Flint or Heath Ledger were. From doing some research, we learn that the reason these two people passed away are related to drug or alcohol overdose. Since the model doesn't have this knowledge, it's not a QA system after all, this example would be really hard for the model to get correct.

Text	**Predicted Label**	**Actual Label**
"When severely depressed I don't feel like I have a disease, I feel that I am a disease. There is nothing to me but pain."	Depression	Bipolar
"I have an addiction to popcorn and Nutella (not at same time) I eat it almost every day like 6/7 in a week"	Anxiety	Addiction
"Keith Flint or Heath Ledger. They both seemed like such bright likable talented souls. Everyone admired them and they had achieved so much."	ADHD	Addiction

Table 5: Misclassified Text Output from Model 2

We also sampled some wrongly classified text for Model 4, which can be found in Table 6. In the first example, the model predicts "depression" even though the actual label is "none." One reason the model might have predicted depression is from the general negative tone of the text and words like “sucks” or “feelings,” which may have tripped the model up. In the second example, the model predicts depression though the actual label is anxiety. This example shows one shortcoming of the way we built our dataset. In particular, since we used pseudo-labels, each post can only have one condition associated with in. In reality, there are people who have been diagnosed with multiple mental health conditions, and our data and thus model are unable to capture those situations. Thus, in this example, we see that the poster has both anxiety and depression, so technically the model’s predictions aren’t incorrect. The last example is an example clearly from r/jokes. The model predicts the label of anxiety. Though unclear exactly why the model predicted so, one potential reason might be the "scare" in "scarecrow." Another reason is that people with anxiety experience excessive worry by definition ([9], [10]), and that worry may sometimes manifest itself through constant questioning.

Text	**Predicted Label**	**Actual Label**
“We suck at conveying feelings. Tell us we are good men once in a while.”	Depression	None
"My kidney surgeon uncle didn’t believe that anxiety and depression were real. He was the worst..."	Depression	Anxiety
“Why did the scarecrow win an award? Because he was outstanding in his field.”	Anxiety	None

Table 6: Misclassified Text Output from Model 4

5.2 Trend Analysis

Time Period	Depression	Anxiety	Bipolar	Addiction	ADHD	None
1.1.20-3.31.20	246	149	10	128	79	2238
4.1.20-6.30.20	240	132	11	121	94	2248
7.1.20-9.30.20	261	112	8	101	91	2272
10.1.20-12.31.20	227	147	7	114	90	2290
1.1.21-3.31.21	221	130	18	132	101	2272
4.1.21-6.30.21	266	138	21	120	83	2244
7.1.21-9.30.21	271	122	9	136	77	2258
10.1.21-12.31.21	298	149	10	123	78	2229

Table 7: Analysis Prediction Counts

To wrap up our project, we did some trend analysis on our analysis data (described at the end of Section 4.1). We started by counting the predictions of the 6 conditions, which can be found in Table 7. As expected, the overwhelming majority of posts are labelled "None." However, we have enough results to find interesting trends in the disorder classes. For example, looking at Figure 3, we see that Depression is the most common mental health condition, regardless of the time period. We also see that Anxiety and Addiction both occur with roughly equal frequency, while ADHD trails slightly behind. Additionally, we see that our model rarely predicted Bipolar Disorder. While this could be an indication that Bipolar Disorder manifests online less frequently, it could also be a failure on the part of our model, as would be suggested during evaluation. Given the explanations in Section 5.1, we see that some of the posts labelled as depression may in reality be Bipolar condition due to the similarity in some symptoms and lack of context provided by the post text.

Next, we looked at how the frequency of mental health disorders changed over the 8 time periods in Figure 4. We see that, overall, the frequency of Depression posts have increased since the start of the pandemic, although it dipped for about six months in the middle. The dip could be explained by the introduction of COVID-19 vaccines, as their roll-out was first discussed at the end of 2020, and many people may have found hope through the news of the roll-out. For a similar reasoning, one may think that anxiety would see a similar dip during this time period. However, we see that anxiety increased towards the end of 2020. This may point at anxiety over the vaccine and its side-effects, which was quite common across the world, though there must have also been people who felt relieved

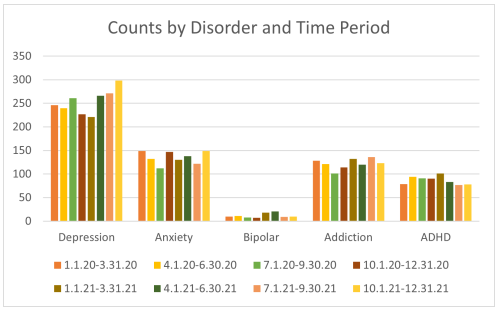

Figure 3: Counts by Condition and Time Period

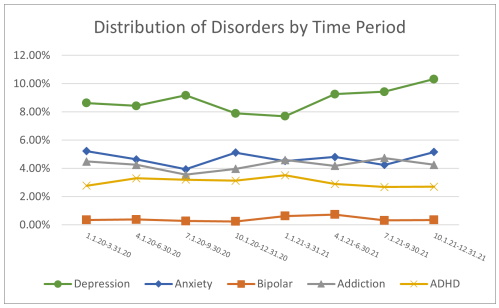

Figure 4: Distribution of Condition by Time Period

that the vaccine would provide further protection from COVID-19. When thinking about this spike, we noticed that the rate of Anxiety also increased during the same months in 2021. Thus, we believe that the spike towards the end of both years for Anxiety could also be caused by holiday-related events and other conflating factors, in addition to increases in COVID-19 cases during this holiday season. Finally, we see that the rates of ADHD and Bipolar Disorder did not meaningfully change over the course of the pandemic. This is likely because ADHD and Bipolar Disorder are both less directly induced by the environment.

We also looked at the distribution of mental health conditions in each of the three subreddits. These trends are shown in Figures 9, 10, and 11. The subreddit with the least fluctuation was r/confessions. Although the rate of Depression was much higher in r/confessions than in the other two subreddits, the frequency of all conditions remained much more stable. When we look at the other two subreddits, we see huge fluctuations. In r/personalfinance, this could be explained simply by an overwhelming lack of posts indicating mental health conditions (over 95% of all posts in r/personalfinance were labeled "None"). Thus, large spikes and dips can be seen from very few changes in classification in terms of counts of people. However, r/legaladvice saw the rates of all conditions (except Depression) fall drastically in the third time period (7.1.20 to 9.30.20). This could be because people were forced to stay home more during this time, so there were less legal concerns in general.

At last, we can look at Figure 12 to see how the predictions are distributed across each of the three subreddits. Again, this shows how r/personalfinance had very few classifications relative to the others. However, it also demonstrates another interesting trend with r/legaladvice in that posts indicating Addiction are much more common. This makes sense, given the topic of the subreddit. r/legaladvice also represents almost half of all the posts indicating Bipolar Disorder. We are not sure why this might be. Perhaps the posts in r/legaladvice tend to express more extreme emotion and the model picked up on that.

6 Conclusion

Overall, we found that classifying mental health conditions is a challenging task. Reddit posts are often laced with sarcasm and unusual vocabulary that make it difficult to infer real meaning, especially regarding the poster's mental health. However, we feel that our attempts to imbue our models with an understanding of emotions and how they relate to mental health ultimately paid off, with our best model achieving an F1 score of over 0.7 for most classes. In the end, we were able to draw interesting conclusions about how certain periods of COVID have affected people's online presence differently. We found that mental health conditions more closely tied to the environment a person lives in were more subject to variation over time. We also found that Depression and Anxiety saw the most potential ties to the COVID-19 pandemic. Finally, we showed that r/legaladvice has higher rates of Addiction and Bipolar Disorder than do the other two subreddits.

In the future, we would likely approach our data differently. Instead of using a randomly selected set of posts from particular subreddits, we would try tracking users we know have mental health disorder for a period of time. This method will better support cases where users are diagnosed with multiple conditions. We would also take the time to experiment with more complex model architectures.

References

[1] Glen Coppersmith, Mark Dredze, and Craig Harman. Quantifying mental health signals in twitter. In *Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality*, pages 51–60, 2014.

[2] Rashmi Patel, Fabrizio Smeraldi, Maryam Abdollahyan, Jessica Irving, and Conrad Bessant. Analysis of mental and physical disorders associated with covid-19 in online health forums: a natural language processing study. *BMJ open*, 11(11):e056601, 2021.

[3] Daniel M Low, Laurie Rumker, Tanya Talkar, John Torous, Guillermo Cecchi, and Satrajit S Ghosh. Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on reddit during covid-19: Observational study. *Journal of medical Internet research*, 22(10):e22635, 2020.

[4] *Diagnostic and statistical manual of mental disorders: DSM-5*. American Psychiatric Association, 2017.

[5] Daniel M. Low, Laurie Rumker, Tanya Talker, John Torous, Guillermo Cecchi, and Satrajit S. Ghosh. Reddit mental health dataset, July 2020.

[6] Pmaw. *PyPI*.

[7] Saif M. Mohammad and Peter D. Turney. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465, 2013.

[8] Philip B Mitchell and Gin S Malhi. Bipolar depression: phenomenological overview and clinical characteristics. *Bipolar Disorders*, 6(6):530–539, 2004.

[9] Samantha Barton, Charlotta Karner, Fatima Salih, David S Baldwin, and Steven J Edwards. Clinical effectiveness of interventions for treatment-resistant anxiety in older people; a systematic review. *Health Technology Assessment*, 18(50):1–60, 2014.

[10] Amy B Locke, Nell Kirst, and Cameron G Shultz. Diagnosis and management of generalized anxiety disorder and panic disorder in adults. *American family physician*, 91(9):617–624, 2015.

Appendix

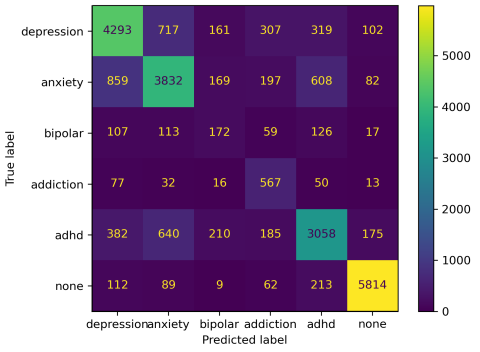

Figure 5: Experiment 1 Confusion Matrix

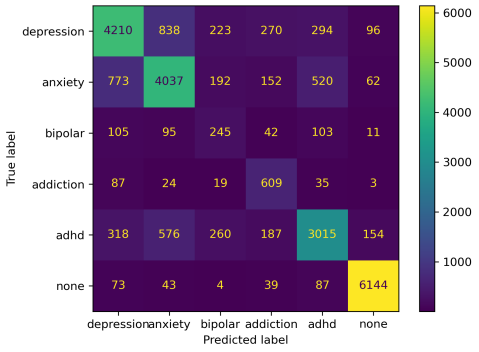

Figure 6: Experiment 2 Confusion Matrix

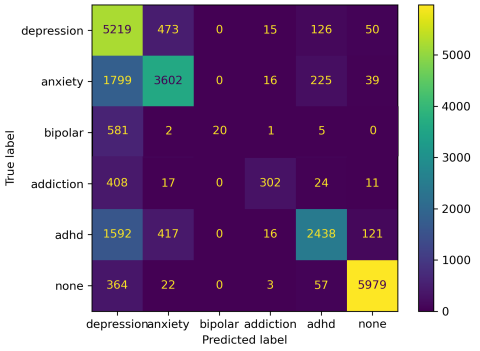

Figure 7: Experiment 3 Confusion Matrix

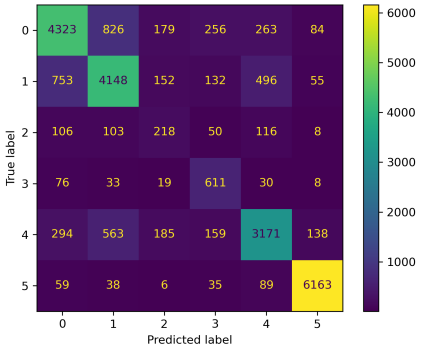

Figure 8: Experiment 4 Confusion Matrix

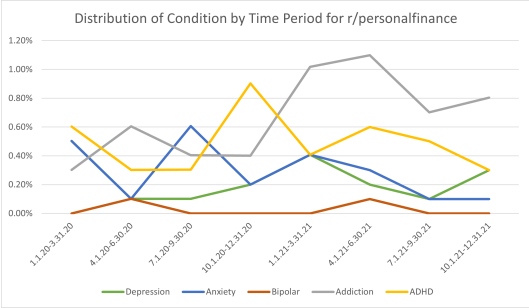

Figure 9: Condition Distribution for r/personalfinance

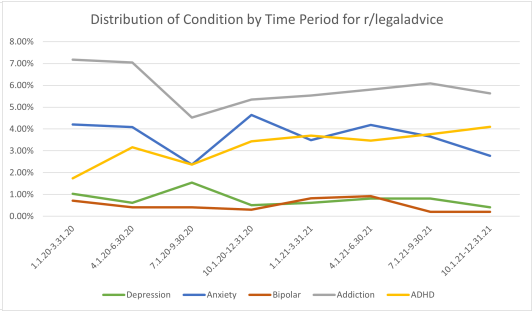

Figure 10: Condition Distribution for r/legaladvice

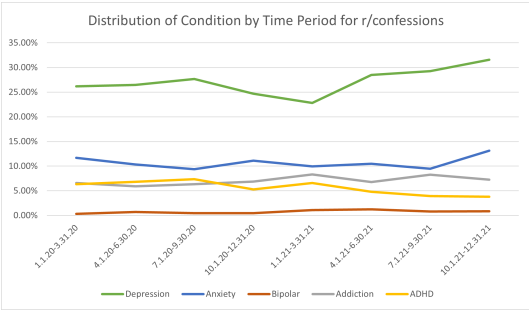

Figure 11: Condition Distribution for r/confessions

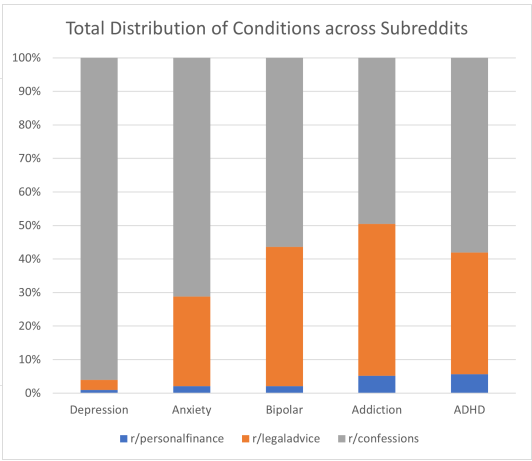

Figure 12: Condition Distributions Across Subreddits