Classification of mental illnesses on social media using RoBERTa

Ankit Murarka[*]
IBM / Raleigh, NC
ankit.murarka1@ibm.com

Balaji Radhakrishnan[*]
balag59@gmail.com

Sushma Ravichandran[*]
IBM Research / Yorktown heights, NY
sushma.ravichandran@ibm.com

Abstract

Given the current social distancing regulations across the world, social media has become the primary mode of communication for most people. This has isolated millions suffering from mental illnesses who are unable to receive assistance in person. They have increasingly turned to online platforms to express themselves and to look for guidance in dealing with their illnesses. Keeping this in mind, we propose a solution to classify mental illness posts on social media thereby enabling users to seek appropriate help. In this work, we classify five prominent kinds of mental illnesses- depression, anxiety, bipolar disorder, ADHD and PTSD by analyzing unstructured user data on Reddit. In addition, we share a new high-quality dataset[1] to drive research on this topic. The dataset consists of the title and post texts from 17159 posts and 13 subreddits- each associated with one of the five mental illnesses listed above or a `None` class indicating the absence of any mental illness. Our model is trained on Reddit data but is easily extensible to other social media platforms as well as demonstrated in our results.We believe that our work is the first multi-class model that uses a Transformer Vaswani et al. (2017)-based architecture such as RoBERTa Liu et al. (2019) to analyze people's emotions and psychology. We also demonstrate how we stress-test our model using behavioral testing. Our dataset is publicly available and we encourage researchers to utilize this to advance research in this arena. We hope that this work contributes to the public health system by automating some of the detection process and alerting relevant authorities about users that need immediate help.

1 Introduction

During these unprecedented times when the world is plagued by COVID19, a large number of people have been showing symptoms of clinical anxiety or depression[2]. This can be attributed to a myriad of reasons including lock down, mandatory social distancing, higher unemployment, economic depression and work-related stress.

In a report published earlier this year, the American Foundation for Suicide Prevention found that people experience anxiety (53%) and sadness (51%) more often now than before the coronavirus pandemic.

In the past decade, social media has transformed how people interact with each other. Apart from sharing factual information and news, people actively partake in sharing their day to day activities, experiences, feelings, opinions, hopes, desires, and emotions online. These texts provide information which can be used to identify the mental health individuals. Furthermore, the current state of enforced social distancing and isolation has propelled more people to express their emotions on social media as it provides them with an accessible platform to share their thoughts with others, many a times, in search for help.

Our research work utilizes user data, especially the kind pertaining to emotions as this class of data can give us valuable insights about the mental state of a person; and, in turn, our work has the potential to assist in the diagnosis and analysis of various mental disorders. This study aims to bridge the gap between people in search of help and experts who can provide the needed help.

Due to the paucity of adequate annotated and structured user data in this domain, we decided to generate our own dataset by crawling subreddits

[*] These authors contributed equally

[1] https://github.com/amurark/mental-health-detection

[2] https://afsp.org

Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis, pages 59–68
April 19, 2021.

on reddit.com[3] pertaining to our use case as a lot of users were found to have shared their feelings there. Although Reddit was the sole source of our dataset, we believe that this study can be seamlessly extended to other social media platforms as well because of the presence of similar unstructured user data online.

The advent of transformers and BERT Devlin et al. (2018) has caused quite a stir in the NLP community because of the state-of-the-art results it was able to produce in various NLP tasks. In this work, we use a RoBERTa Liu et al. (2019) based classifier, which has a similar architecture to BERT with an improved pre-training procedure. RoBERTa's effective and efficient performance on unstructured data and its ability to learn contextual information compelled us to explore and capitalize its power to categorize online user generated texts into various classes of Mental Illness.

We identify five broad classes of mental illnesses - depression, anxiety, bipolar disorder, ADHD (Attention Deficit Hyperactivity Disorder), PTSD (Post Traumatic Stress Disorder) and an additional 'None' class (which does not pertain to any mental illness). We train a multi-class classifier on the data crawled from online user data. Based on our experiments, we present encouraging results that demonstrate that social media data has the potential to complement standard clinical procedures in the prognosis of mental health amongst two broad categories of users - ones who are seeking help online and ones who are unbeknownst of their condition.

2 Related Work

In the recent past, people have increasingly turned to social media to share and seek counsel on the topic of mental health. This has prompted researchers to utilize the information and apply a plethora of techniques in NLP and Machine Learning in order to assist people who might require help. Most of the recent such research, as described in the subsequent paragraphs, has revolved around Reddit data: Kim et al. (2020), Gkotsis et al. (2017), Sekulic and Strube (2019), Zirikly et al. (2019). Prior to this recent shift to Reddit data, a lot of the earlier research was focused on utilizing Twitter data: Orabi et al. (2018), Benton et al. (2017), Coppersmith et al. (2015).

There have been a wide variety of approaches ranging from classical NLP techniques to neural network based deep learning methods. Coppersmith et al. (2015) used character level language models to examine how likely a sequence of characters is to be generated by a user with mental health issues. Benton et al. (2017) evaluated a standard regression model, a multilayer perceptron single-task learning (STL) model, and a neural MTL model on detecting multiple types of mental health issues. Orabi et al. (2018) utilized word embeddings in tandem with a variety of neural network models like CNNs and RNNs to detect depression. Gkotsis et al. (2017) experimented with Feed Forward Neural Networks, CNNs, SVMs and Linear classifiers to perform binary classification on mental health posts. Sekulic and Strube (2019) came up with the approach of using Hierarchical Attention Networks(HANs) to detect a wide range of mental health issues like Depression, ADHD, Anxiety etc. and trained a binary classifier for each of the disorders. The most recent work on this was by Kim et al. (2020) who proposed a CNN-based classification model. Once again though, each disorder had its own separate binary classifier to perform the detection.

To our knowledge, this is the first attempt at treating this problem as a multi-class classification problem, where a single classifier can accurately classify the type of disorder that the person is referring to in their post. In addition, this is also the first work that harnesses the incredible capabilities of an advanced Transformer based algorithms like RoBERTa to solve this difficult problem.

3 Dataset

The Reddit API was used to crawl 13 Reddit Subreddits for a total of 17159 posts (text and title) to obtain the data for this work. The text from comment threads of these posts was not collected as it tended to diverge from the main topic of the subreddit. Even though there are a lot of mental illnesses that need addressing, only 5 of them had sufficient data for our purposes that were chosen for the purposes of this paper. They are: `bipolar`, `adhd`, `anxiety`, `depression` and `ptsd`. We do plan to extend this solution to tackle the remaining illnesses as well in due time. The posts in these subreddits were assigned a class label corresponding to the name of the mental illness they were associated with. All the remaining subreddits were carefully chosen as to minimize any chances of thematic content overlap between them and the illness classes

[3] https://www.reddit.com

i.e to enable the model to differentiate between posts discussing mental illness from those that were not, we chose a few subreddits from a wide range of common topics such as `music`, `travel`, `india`, `politics`, `english`, `datasets`, `mathematics` and `science`. The text from these subreddits were combined and assigned the class label None. While we are currently treating this task as a multi-class classification problem, we duly acknowledge the fact that mental illnesses are generally co-related and require multi-label classification techniques. We are actively working on converting this task to a multi-label classification problem.

While collecting data, we ensured that the number of upvotes for each post in all subreddits is more than 10. We also set a minimum post token length of 30 tokens. These numbers were chosen after carefully perusing through the data and going over the post text in order to retain quality in the dataset. We initially crawled the subreddits pertaining to mental illness and were able to collect about 3000 posts in total. We then crawled the other subreddits corresponding to the None class label and collected about 300 posts per subreddit for 8 subreddits. This also ensured a good balance of class labels. While selecting the eight general topic subreddits, we not only selected subreddits that have sufficiently high number of posts, but also ensured that we cover a broad range of topics. Table 1 shows the statistics collected for each subreddit. The dataset was preprocessed to remove any URLs or usernames that could potentially contain sensitive information. This was done keeping in mind that the dataset will be released publicly for the purpose of extending this research work.

To gauge the data quality we ran some analysis. We manually went over the lowest voted posts for each mental illness subreddit. We wanted to establish that texts from these posts expressed emotions from people discussing corresponding mental illness it is labelled as. Table 2 presents excerpts of lowest voted post from each mental illness subreddit.

We also certified that the general topic subreddits did not have a high similarity with the posts corresponding to other the other 5 subreddits. This was done to ensure that we do not have any false negatives while assigning truth labels. We counted the number of posts the mental illness terms appeared in, for each subreddit. The subreddits corresponding to mental illnesses had a much higher count of these words. In addition to this, we compared the cosine similarity between some of the highest/lowest posts of mental illness subreddits and the general topic subreddits and manually compared the results to find that this distance was higher than the distance between two posts of the mental illness subreddits.

We also attempted to augment the data using Easy Data Augmentation Wei and Zou (2019) to boost the performance of our model. However, we did not observe an apparent shift in our evaluation metrics- explained by the fact the EDA is meant to perform best for smaller datasets ($\leq$ 5000 sample sizes).

4 Model

In this section, we describe our model architecture for the multi-class mental illness classification task. We propose a RoBERTa based classifier in order to accomplish this. In addition, we also compare the proposed model against an LSTM Hochreiter and Schmidhuber (1997) based classifier and a BERT Devlin et al. (2018) based classifier to demonstrate the superiority of our approach. Since this is an entirely new dataset, there is no established baseline, so the LSTM model will serve as the baseline for our experiments. We also showcase our gains over BERT, the most widely used transformer model today for text classification. All our models were implented in Pytorch Paszke et al. (2019). The Transformer models were implemented with the help of the HuggingFace Transformers Wolf et al. (2019) library.

4.1 LSTM based classifier

LSTMs(Long Short-Term Memory) were the state of the art models when it came to text classification before the advent of Transformers. They will serve as our baseline. First we tokenized the sentences using NLTK[4] and converted them to lower case to create our vocabulary. In order to get rid of words that might not exist, we removed all words from our vocabulary that appear only once. We also added `padding` and `unknown` to our vocabulary in order to account for padding and unknown tokens respectively. Each sentence was represented using a sequence of length 512 and this forms our input to the LSTM model. We used a 2 layer LSTM for all our experiments with an embedding layer of size 100 and a hidden layer size of 256. Dropout

[4] https://www.nltk.org/

Subreddit	Number of posts	Average no. of words (posts)	Average no. of words (titles)	Average Upvotes	Highest Upvotes	Lowest Upvotes
r/depression	3062	152.74	12.20	517.19	4802	11
r/anxiety	3027	170.38	11.75	246.07	3349	11
r/ptsd	2501	233.55	10.14	38.4	443	11
r/adhd	3082	198.55	13.71	377.13	4484	11
r/bipolar	3009	203.28	9.26	32.37	363	11
none	2478	238.52	15.76	6715.33	199295	11

Table 1: Dataset: Statistics

Lowest rated post
r/depression- The older I am getting the less hope I have to secure a life worth living. I feel finished because I always had that state of mind **r/adhd**- Does anyone else feel like u do have a personality and ability to make friends but ure kind of stuck in ur own body **r/bipolar**- I just made a really impulsive choice with my breed of dog because I had to have one NOW. I was thinking about it constantly day and night and I couldn't sleep. **r/anxiety**- I find myself constantly remembering embarrassing or cringey moments from my past (ranging anywhere from present day to back about 10 years) and cringing hard at them **r/ptsd**- I feel like I am just constantly angry. Angry about my trauma and how it has affected me, and angry about where I am in my life because of it. I don't want to be angry anymore

Table 2: Dataset: Posts excerpts

Srivastava et al. (2014) with a probability of 0.5 was used in order to achieve regularization. We used standard cross-entropy loss as the loss function. During training, Adam Kingma and Ba (2014) was the optimizer of choice, with a learning rate of 0.005. The model was trained for a total of 25 epochs with a batch size of 32. Gradient clipping was used to prevent exploding gradients.

4.2 BERT based classifier

BERT(Bidirectional Encoder Representations from Transformers) has been the biggest breakthrough in NLP in the recent past with state of the art results in a myriad of tasks. Since its inception, better models with gains have been trickling along, but BERT continues to be the most popular model for text classification even today. The BERT classifier comprises a fine-tuned BERT model followed by a dropout layer and a fully connected layer. We fine-tuned a pre-trained BERT-base model on our dataset for this task. A pre-trained tokenizer on BERT is used to tokenize our input sentences. After carefully examining the sentence length distribution, we chose a sequence length of 35 for titles, and 512 for posts and posts+titles. Either padding or truncation was used to ensure that all sentences were represented using the same sequence length. All the BERT based models were fine-tuned on our data for 10 epochs with a learning rate of 1e-5. Adam served as the optimizer and cross-entropy loss was the loss function of choice. A dropout layer with probability of 0.3 was used for the sake of regularization.

4.3 RoBERTa based classifier

RoBERTa(Robustly Optimized BERT Pretraining Approach) is another state of the art language model that builds on BERT by modifying key hyperparameters and training on more data. It outperforms BERT on several benchmark tasks and forms the core of our proposed solution. In order to make it a fair comparison with BERT, we retain the architecture and all design choices made with the BERT based classifier barring the pretrained model and the tokenizer which are now all based on RoBERTa. The input sentences were tokenized using a pre-trained tokenizer on RoBERTa-base. Just as in the case of BERT, we chose a sequence length of 35 for titles, and 512 for posts and posts+titles. Similar to BERT, the RoBERTa based models were also fine-tuned for 10 epochs with a learning rate of 1e-5 and Adam. A batch size of 32 was used while fine-tuning on the titles whereas a batch size of 16 was the only viable option to fine-tune on posts and posts+titles. Cross-entropy remained the preferred loss function. Once again, a dropout layer with probability of 0.3 was used for regularization.

5 Result Analysis

As described earlier, in addition to our primary RoBERTa classifier, we also run experiments on an LSTM classifier and a BERT classifier for the sake of comparison. We fine-tune each of the aforementioned models on just the titles, just the posts and a combination of both in order to perform comprehensive tests and comparisons. When combining the titles and posts for our Transformer models, we convert the problem into a sequence-pair classification task. This allows the model to give more importance to the title which would otherwise be lost when combining the title and the post into one single input given the relative difference in their lengths (the average number of tokens in titles is roughly 6% that of posts).

The results from our experiments are documented in Tables 3 through 5.

As can be observed from Table 3, our proposed RoBERTa based classifier far outperforms the baseline LSTM in all categories. The BERT classifier has results which are quite close to that of RoBERTa's and both beat LSTM by a significant margin, showcasing the incredible capabilities of pre-trained Transformer based architectures. In fact, our RoBERTa model fine-tuned on just the titles was able match the performance of the LSTM model trained on posts. The RoBERTa model was able to achieve an F1 score of 0.86 on the posts and 0.89 on posts+titles which are extremely promising given the complex nature of the multi-class mental illness classification task. The jump in accuracy between posts and posts+titles is not as drastic as the jump between titles and posts. This indicates that the posts offer far more valuable information when compared to the titles and also the fact that most of the useful and relevant information can be extracted from the posts alone. This strong performance on just the posts bodes well for the extensibility of our approach as this can be applied on almost any given social media post without the need for structure in the data like titles, user names, user history, etc.

The rest of this section will focus solely on the results of our best performing RoBERTa model. Table 4 showcases the granular class-wise results of the RoBERTa model. This table in conjunction with the confusion matrices from Figure 1 offers us a wealth of useful and interpretable information.

The first strikingly obvious result is the high accuracy with which the model is able to detect non-illness related posts. Even with just the titles, the model is able to classify the `none` class with an f1 score of more than 0.9. This gives us hope that this model will suffer from very few false positives when it comes to mental illness detection on social media.

An even more crucial property of our model can be noticed in the confusion matrices for posts and posts+titles in Figure 1. When using posts, just 3 illness related posts across the entire test dataset were misclassified as non-illness posts. This number further reduces to 0 when using titles+posts. This shows that the model will detect mental illness posts correctly nearly every single time, thus ensuring that posts from people who are seeking help never go unnoticed when this solution is deployed in the real world.

When it comes to the class wise performance amongst the mental illnesses, the two best performing classes are `adhd` and `ptsd` whereas the two worst performing classes are `depression` and `anxiety`.

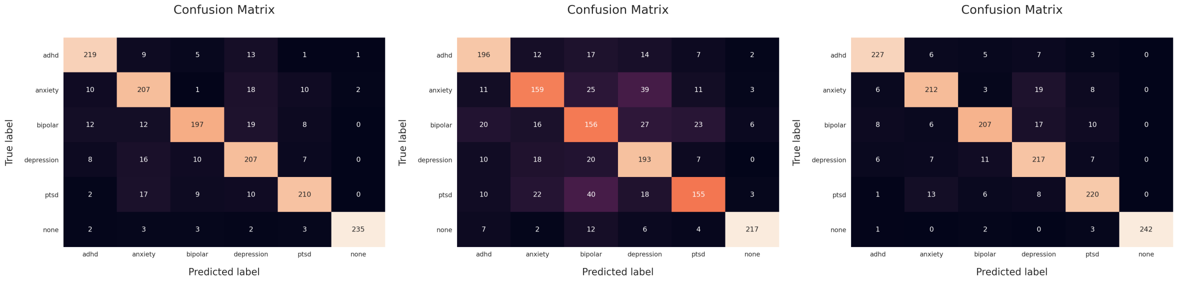
The performance of `depression` and `anxiety` classes can be attributed to a few factors. The average number of words per post for `depression` and `anxiety` are the least for any given class. For instance, `depression` posts have roughly 53% lesser textual data when compared to `ptsd` posts. In addition, studies show that depression might often occur in tandem with another mental illness and our data and results back this up as well. The `depression` word occurs in 12% of `anxiety` posts, 12% of `ptsd` posts and 31% of `bipolar` posts. Similarly, `anxiety` occurs in 20% of `ptsd` posts, 12% of `adhd` posts and 14% of `bipolar` posts. This implies that the model cannot give high importance to the mention of these class names like it can with rest of the illnesses, thus making the classification of these 2 classes that much harder.

This can also explain the relatively lower precision scores(higher number of False Positives) for `depression` and `anxiety`. When the other illnesses(excluding `depression` and `anxiety`) are misclassified, they are almost always misclassified as either `depression` or `anxiety`, as can be viewed

[5] In all our tables, P and R stand for Precision and Recall respectively

Models	posts				titles				posts+titles			
	P	R	F1	Acc	P	R	F1	Acc	P	R	F1	Acc
LSTM	0.74	0.72	0.72	0.72	0.65	0.64	0.64	0.64	0.77	0.76	0.76	0.76
BERT	0.83	0.82	0.82	0.82	0.72	0.71	0.71	0.71	0.87	0.87	0.87	0.87
RoBERTa	0.86	0.86	**0.86**	0.86	0.73	0.72	**0.72**	0.72	0.89	0.89	**0.89**	0.89

Table 3: Results: Classification Report



(a) Input: posts

(b) Input: titles

(c) Input: posts+titles

Figure 1: RoBERTa: Confusion Matrices

Class	posts			titles			posts+titles		
	P	R	F1	P	R	F1	P	R	F1
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

Table 4: Results: RoBERTa Class-wise results

in Figure 1. In the same figure, we can see that `depression` and `anxiety` are often misclassified as each other due to the reason that they commonly occur together.

There are more posts in the `adhd` and `ptsd` classes that mention the words `depression` and `anxiety` than their respective class names itself. One would assume that this would result in sub-par results, but, these classes actually perform the best. This really showcases the true potential of our model, where it doesn't just rely on mention of class names, but has a strong understanding of the context of the post itself. Additionally, the symptoms or descriptions provided for these classes could be strong, unique and discriminative enough for the model to be able to classify them correctly even with all the mentions of other class names.

In Table 5 we have documented a few interesting results we observed in the test set. In the first two examples on the table, the RoBERTa model was able to classify the posts correctly without the presence of class names in the input. The prediction is based purely on contextual information learnt about the class labels during the training process. The next two results are interesting because the truth label assigned to the input text may or may not correspond to actual mental illness described in the text. Since, we are not domain experts ourselves, we would need expert intervention to substantiate this theory. As a part of future work, getting professionals to annotate our dataset might help strengthen the model for such examples. Further samples from our qualitative testing on various social media platforms can be found on our accompanying web-page[6].

Input	Actual	Predicted
often times i'll get distracted from my thoughts either by external influences or just another idea coming in, and then i have to spend a good 5 minutes trying to work out what i was thinking about again.	adhd	adhd
once i come down from flashbacks or panic attacks, i get really bad disassociation. sometimes lasting for days. does anyone else go through this. any tips on how to stop it. i tried grounding but im so far gone it doesn't help.	ptsd	ptsd
i can't sit still when i get my eyebrows done, and when i'm in class i usually doodle to focus. i pay attention very well in school regardless of that, and drawing helps me focus.	anxiety	adhd
i'm flying from dallas to hong kong in january and it's 17 hours. i've flown 12-13 hour flights before and they really mess with me. so i'm wondering - what are your tips for not going crazy on such a long flight? ps: i'm terrible at sleeping on planes. thinking about taking some sleepy meds to see if it'll help	none	anxiety

Table 5: Results: Interesting Examples

Synonym Replacement					**Label Removal**				
Test Set Modified	**posts**				**Test Set Modified**	**posts**			
	P	R	F1	Acc		P	R	F1	Acc
10%	0.86	0.85	0.85	0.85	10%	0.85	0.84	0.84	0.84
50%	0.85	0.84	0.84	0.84	50%	0.81	0.80	0.80	0.80
100%	0.83	0.83	0.83	0.83	100%	0.75	0.74	0.75	0.74
Test Set Modified	**titles**				**Test Set Modified**	**titles**			
	P	R	F1	Acc		P	R	F1	Acc
10%	0.73	0.72	0.72	0.72	10%	0.72	0.71	0.71	0.71
50%	0.71	0.71	0.71	0.71	50%	0.67	0.67	0.67	0.67
100%	0.68	0.67	0.67	0.67	100%	0.61	0.61	0.60	0.61
Label Replace: 'illness'					**Label Replace: random**				
Test Set Modified	**posts**				**Test Set Modified**	**posts**			
	P	R	F1	Acc		P	R	F1	Acc
10%	0.84	0.83	0.84	0.83	10%	0.83	0.82	0.83	0.8
50%	0.78	0.77	0.77	0.77	50%	0.71	0.71	0.71	0.71
100%	0.70	0.67	0.68	0.67	100%	0.58	0.57	0.57	0.57
Test Set Modified	**titles**				**Test Set Modified**	**titles**			
	P	R	F1	Acc		P	R	F1	Acc
10%	0.72	0.71	0.71	0.71	10%	0.71	0.71	0.71	0.71
50%	0.67	0.65	0.65	0.65	50%	0.64	0.64	0.64	0.64
100%	0.62	0.57	0.58	0.57	100%	0.54	0.54	0.54	0.54

Table 6: Behavioral Tests

6 Behavioral Testing

Although the classification metrics analyzed in the previous section are generally regarded sufficient in estimating the performance of Bert-based models, a recent inclination of NLP researchers to behavioral testing inspired us to stress test our models as well.

For all our tests, we used our proposed RoBERTa model and applied these tests to inputs that were either titles or posts. Since we hope to extend our model to other social media platforms, we do not always expect input texts to have a title as well as a descriptive text/post. We adopted the Checklist approach Ribeiro et al. (2020) which involve tests conducted to comprehensively analyze the model's performance.

6.1 Synonym Replacement

Synonym replacement is a kind of Invariance Test where label-preserving perturbations are made to the test set. As labels, the root form of the mental illnesses was chosen- `depress`, `ptsd`, `anxiou/anxiet`, `bipolar` and `adhd`. Python's NLTK package and WordNet were used for these tests.

This test is conducted such that the root words are not perturbed when modifying the test set. For each post, 10% of the tokens were randomly selected (not including the stop words or the root words). Each token was then replaced with one of its synonyms. We used the same logic for titles. We set a max and min on the number of tokens to be selected for replacement - this was (4, 30) for posts and (1, 5) for titles. Since each token was replaced with a synonym, the class label for the samples was not changed. We did this for 10, 50 and 100 percent of the test set and observed results.

In all three cases, we expect the classification metrics to drop. For the case when 10% of the test case was modified the drop was much lower as compared to when 100% of the test case was modified. The results are documented in Table 6. When comparing these results to those in Table 4 we find that the drop in each category is about 2-4% for posts and 5-7% for titles. The lower drop can be attributed to the fact that synonym replacement does not alter the semantics of the input text. Therefore the model was able to draw sufficient information from the input.

6.2 Masking

We also performed a Directional Expectation test on the model. This is similar to the previous test but is instead performed only on labels. The labels, as defined in the previous subsection, are a list of the root form of mental illness class labels. We noticed that the root words appear often in our input texts. This behavioral test was performed to observe our model's dependency on these words. For all the tests below, we modified only those tokens that contained a root word.

In the first case, for every post from a subreddit related to a mental illness, the root form of its class label was removed from the input. For example, the input text: *I feel happy for some time and then depressed again. I'm definitely bipolar* from the `r/bipolar` subreddit, was modified to *I feel happy for some time and then depressed again. I'm definitely*. Note that changes were not made to the word *depressed* in the input. The class label for each modified sample was not changed after the perturbations. Like the previous subsection, these tests were performed on 10, 50 and 100% of the test set.

In the second case, instead of entirely removing the tokens, we replaced it with a generic token `illness`. We expected this modification to retain some semantic information that was lost in the previous test. However, we found that adding a generic token introduced some noise which reduced the overall performance of the model.

Lastly, the tokens were replaced by a randomly chosen root form of a mental illness other than its class label. With this test, we expect to force the model to pick between the label and non label tokens during classification. We believe that this is an interesting scenario to observe.

In all three cases (Table 6), the model performance drops by some degree when compared to Table 4. The first two cases showed a somewhat similar performance drop. However, the model performance was worse than that of the Synonym Replacement test. This means that the model depends on the existence of the root words in the input text to some degree.

In the third scenario, we note that the performance drop is higher. Although the test is meant to confuse the model, we observed that in some cases (especially for input: posts), we got an F1 of 0.82 with 10% of the modified test and 0.71 with 50% of the modified test. This is only possible if the

[6] https://mental-health-classification.github.io/

model gathered sufficient information from the non label text in the input.

7 Conclusion and Future Work

Our chief motivation behind this work is the current pandemic and mandatory confinement worldwide. We believe that social media has become the prime mode of communication for people and has paved way for them to vent freely without judgement.

Our roadmap includes getting all our data annotated by mental health experts in order to verify our annotations. This would also assist us in creating a multi label dataset which is more representative of this problem when compared to a multi class one. We would also like to work on bettering our model on the behavioral tests. Our work involves two kinds of texts- long and short - both of which are common to the internet community. Hence, our work can easily be extended to many websites.

In conclusion, we believe that our work explores an interesting line of research where NLP is used to bridge the gap between virtual and real life of users and help those in need of medical attention.

References

Adrian Benton, Margaret Mitchell, and Dirk Hovy. 2017. Multi-task learning for mental health using social media text.

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 2015. Clpsych 2015 shared task: Depression and ptsd on twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

George Gkotsis, Anika Oellrich, Sumithra Velupillai, Maria Liakata, Tim J. P. Hubbard, Richard J. B. Dobson, and Rina Dutta. 2017. Characterisation of mental health conditions in social media using informed deep learning. *Scientific Reports*, 7(1):1–10.

Sepp Hochreiter and Jurgen Schmidhuber. 1997. Long short-term memory.

Jina Kim, Jieon Lee, Eunil Park, and Jinyoung Han. 2020. A deep learning model for detecting mental illness from user content on social media. *Scientific Reports*, 10(1):1–6.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Ahmed Husseini Orabi, Prasadith Buddhitha, Mahmoud Husseini Orabi, and Diana Inkpen. 2018. Deep learning for depression detection of twitter users. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology*. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of nlp models with checklist. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Ivan Sekulic and Michael Strube. 2019. Adapting deep learning methods for mental health prediction on social media. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 322–327, Hong Kong, China. Association for Computational Linguistics.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Ayah Zirikly, Philip Resnik, Özlem Uzuner, and Kristy Hollingshead. 2019. Clpsych 2019 shared task: Predicting the degree of suicide risk in reddit posts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*. Association for Computational Linguistics.