Classification of mental illnesses on social media using Random Forest Classifier

Jaymar Cezar
CCIT
National University
Manila, Metro Manila, Philippines
cezarjr@students.national-u.edu.ph

Sean Patrick Gabriel
CCIT
National University
Manila, Metro Manila, Philippines
gabrielst@students.national-u.edu.ph

I. Introduction

Our emotional, psychological, and social well-being are all parts of our mental health. It influences our thoughts, emotions, and behaviors. Additionally, it influences how we respond to stress, interact with others, and make decisions. Every period of life, from childhood and adolescence to maturity, is vital for mental health. Mental health disorders such as depression, anxiety, ADHD, etc., are common illnesses that people suffer from. Due to their less noticeable illness, it is difficult to diagnose what kind of mental disorder they suffer from, unlike physical injuries that can be observed by using machines such as X-Ray or CT scans. The only way to diagnose these illnesses is by seeking out help from others, but not all individuals have the courage to seek help rather they keep it to themselves. As a result of barely noticing signs from an individual to detect whether he or she suffers from a mental disorder, the number of suicides continually arise. According to the statistics given by World Health Organization (WHO) in 2021, it is stated that more than 700,000 people die due to suicide every year. This number could be lessened with proper treatment among individuals. The proper diagnosis of patients with mental illnesses is difficult to recognize due to many aspects such as physical attendance, social issues, and not being comfortable expressing thoughts.

In today's society, social media has drastically changed how individuals interact and communicate with one another. People are actively participating in sharing their daily activities, experiences, sentiments, views, hopes, desires, and emotions on social media. Each individual's post contains details that can be utilized to recognize people with mental illnesses.

This is where we got the idea of analyzing and classifying a complex model on social media (Reddit) posts that are at risk of mental illnesses. This focuses on detecting 5 types of mental illnesses which are ADHD, depression, anxiety, bipolar, and PTSD. We used a learning method for classification which is the random forest. According to [1] one of the most popular classifiers for classification and regression problems is random forest (RF). It is a preferred solution for text classification due to its straightforward algorithm. Additionally, it has a substantial advantage over other machine learning models, it is that it can handle high-dimensional data and perform well with unbalanced datasets [27,28,29,30,31]. In RF, a significant number of decision trees are used to make decisions based on the "knowledge of the crowd." In contrast to the decision tree, it produces more accurate results since it takes the average or means of the decision tree outputs to make the conclusion. RF is used for handling non-linear classification tasks.

However, this is not the optimal solution due to the unethical practice of obtaining data from each individual. This project tries to fill the gap between those in need of assistance and the health professionals who can diagnose and treat their mental illnesses.

II. Related Work

In recent years, NLP researchers have demonstrated significant interest in the area at the convergence of Machine Learning and Psychiatry. Social media is a crucial research resource. Many researchers initially concentrated on the Twitter text [2,4,5], but then switched their attention to the Reddit platform [24,25,4,26].

To analyze texts about mental health, a variety of methods have been used, from conventional ML to sophisticated DP. [5] states that they used character-level language models to determine how likely it was that a user with mental health issues would produce a group of characters. [4] used neural MTL, regression, and multi-layer perceptron single-task learning (STL) models to identify several types of mental health issues. The most important step was removing the acronym ADHD from the messages before learning, and further information concerning attention issues was eliminated from the texts. [3] trained the SVMs to distinguish 200 text messages into two classes: "ADHD or not." The objective was to examine how effectively the SVMS learns in the absence of keywords and even semantically pertinent content.

In another study, [7] tested cutting-edge techniques using Reddit's SMHD mental health conditions dataset proposed by [22]. Their three contributions are: focusing on general text rather than mental health support groups; classifying by posts rather than individuals or groups; and using a dataset with more disorders than most research. They use BERT, RoBERTa, and XLNET, three deep learning models for automatically classifying disorders. On a small portion of [22]'s dataset, they double the baseline they set. They enhance [23]'s post-level categorization findings. The presence of discussions about calories, diets, recipes, etc. gave the eating disorder classifier the best accuracy results; depression, on the other hand, had the lowest F1 score, perhaps because it is more challenging to identify depression in linguistic acts.

Data mining techniques like the Random Forest classifier [8] have been found to provide better classification performance than other cutting-edge algorithms [9]. These characteristics have increased the popularity of Random Forest during the past few years, particularly in the study of mental illnesses [10, 11]. The random forest classification approach and the impute missing values learner [12] are integrated in this study's mental illnesses procedure and decision-making process to forecast missing values in the dataset. The study's methodology involved categorizing patients with addictions based on substance misuse and categorizing those with psychoses based on various psychopathologies' characteristics (attributes). The discovery of hidden data that cannot be exposed using conventional brain imaging [13, 14] techniques is made possible by data mining approaches.

III. Methods

A. Dataset

This study uses reddit.com posts data proposed by Murarka and Radhakrishnan [21]. The Reddit post dataset has been designed to classify mental diseases into one of five categories. The dataset contains a total of 16,703 posts. The dataset was further split into a train, dev, and test set, with 13,727 posts in the training set and 1,488 posts in both the test and validation sets. The dataset comprises five columns: id, title, post, 'class name,' and 'class id.' Figure 1 depicts the total number of posts for each mental illness class in the whole dataset.

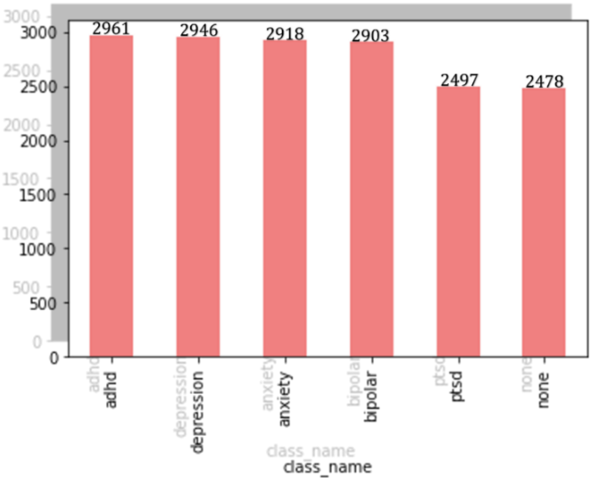


Figure 1. Number of posts for each mental illness class in whole dataset

B. Data Cleaning

Data cleaning is a critical stage in any machine learning project, particularly an NLP one. We begin the data cleaning procedure by removing the unnecessary columns, which are the id and class name columns. We then changed all of the text in the posts and titles to lowercase. URLs and numbers were eliminated. We also removed all of the punctuation marks. non-ASCII characters and stop words or words that occur often in the corpus but are regarded as low-level information. We eliminated those words and characters to make the dataset cleaner and to give more focus on the important information. We also applied lemmatization to reduce the amount of information that the computer had to cope with, and therefore improve efficiency.

C. Features

Feature engineering is the process of extracting relevant information from raw data in order to make it usable for machine learning models. It enables us to generate better data, which allows the model to interpret it and provide reasonable results. Features are essentially the model's inputs.

The first attribute we retrieved is the total amount of words for both the post column and the title column. This is a very common and simple feature extraction technique.

We also extracted the counts of part-of-speech (POS) tags for verbs, nouns, adverbs, and adjectives in the post column. We first applied Parts of Speech (POS) tagging which is a process of marking each word in a corpus with its corresponding part of speech. Then we only count the number of POS tags for verbs, nouns, adverbs, and adjectives since these four are the most important parts of speech, and because we already removed the stop words that are commonly pronouns, conjunctions, interjections, and prepositions. We only extracted the POS tag counts for the posts since most of the texts in the title are very few. However, we still tried extracting the POS tag counts for the titles, but it yielded a lower accuracy.

We also retrieved the text polarity of the posts. The degree of negativity or positivity in a piece of text is measured by its text polarity. Since our dataset consists of social media posts, we employed VADER (Valence Aware Dictionary and sEntiment Reasoner), a lexicon and rule-based sentiment analysis tool that is especially geared to sentiments expressed in social media [15]. We only extracted the polarity of the posts because on the title, it will always just produce 0.0 for compound, positive, and negative and 1.0 for neutral since the texts in the title are very short. We also tried getting the text polarity of the titles, but it resulted in a lesser accuracy.

The number of times a certain mental illness word stated we’re also extracted from both the post and title column. We also extracted the number of occurrences of words that may signal a particular mental illness, words that are related, and symptoms for each mental illness. We created columns for the number of occurrences of words for each mental illness. The words we extracted for depression words are words that may signal depression that we got from an article on thehealthy.com website written by Tina Donvito, a health and wellness writer and blogger [16]. Tina Donvito acquired the words and phrases she wrote in the article from a study that examined online forums and published in Clinical Psychological Science. Some of the words that we extracted for depression are absolute, completely, upset, bad, sad, helpless, aching, lost, worthless, and useless. The words we used for anxiety, ADHD, bipolar, and PTSD are combinations of words that are related to and symptoms of each mental illness. We acquired the words we extracted for those mental illnesses from the website of the Mayo Clinic, a nonprofit academic medical facility in the United States that focuses on integrated health treatment, education, and research [17,18,19,20]. We did not extract count of some words that may be used to describe various mental illnesses like the word sleeping which could both signify anxiety and PTSD when there is a phrase that says, ‘trouble sleeping.' However, some words, like "hyperactivity," "inattention," and "impulsivity," were manually assigned to a particular list of mental illness words depending on what we read that particular words describe. Those words are major symptoms of ADHD, but they can also indicate PTSD, which is why diagnosing PTSD and ADHD can be challenging because the symptoms of PTSD resemble those of ADHD. We categorized the terms hyperactivity, inattention, and impulsivity as words for ADHD since those words are significant signs of ADHD.

D. *TF-IDF*

Using Term Frequency-Inverse Document Frequency (TFIDF), we converted the texts in the posts into vectors and assessed the significance of each word in the corpus. In text mining applications, it serves as a weighting element. By combining a word's Term Frequency (TF) and Inverse Document Frequency (IDF), TF-IDF vectorizes or scores a word.

Term Frequency (TF) measures how frequently a term or word appears in a document compared to the total number of words in the document.

$$TF = \frac{\text{number of times the term appears in the document}}{\text{total number of terms in the document}}$$

Inverse Document Frequency: IDF of a term reflects the proportion of documents in the corpus that contain the term. Words unique to a small percentage of documents (e.g., technical jargon terms) receive higher importance values than words common across all documents (e.g., a, the, and).

$$IDF = log(\frac{\text{number of the documents in the corpus}}{\text{number of documents in the corpus contain the term}})$$

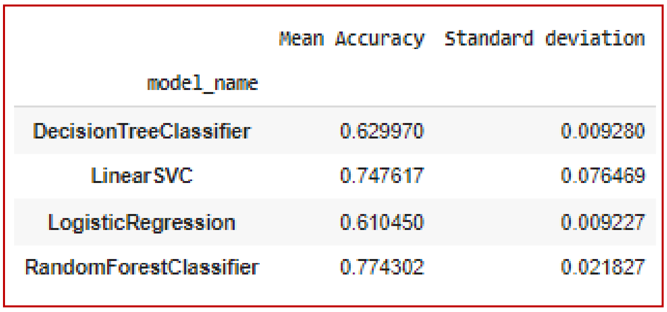
The TF-IDF of a term is calculated by multiplying TF and IDF scores.

$$TF\text{-}IDF = TF * IDF$$

We used the scikit learn implementation of the TfidfVectorizer function with the default parameters with a few changes using these parameters: sublinear_tf=True, min_df=5, stop_words='english'. The min_df is set to 5 meaning it will ignore or remove the words which have occurred in less than 5 documents. It is also set to apply sublinear scaling or logarithmic scale, and it is also set to remove stop words although we already removed the stop words in data cleaning.

E. *Algorithm Selection*

In choosing the algorithm we used, we first combined the training set and validation set before using k-fold cross-validation with five folds to determine the cross_val_score for each algorithm. The scikit-learn package's cross_val_score function trains and evaluates a model or algorithm over several folds of the dataset. We utilized k-fold cross-validation because, despite having a higher computational cost than a single validation set, it prevents hyperparameters from overfitting to a fixed validation set and makes better use of the data by using the complete concatenated training and validation set across the folds. We compared four different machine learning classifiers: Decision Tree, Linear Support Vector Machine, Logistic Regression, and Random Forest Classifier. As we can see in Table 1 below, we can clearly say that the Random Forest Classifier outperformed all the other classification algorithms, and that's why we chose to move forward with the Random Forest Classifier.



	Mean Accuracy	Standard deviation
model_name		
DecisionTreeClassifier	0.629970	0.009280
LinearSVC	0.747617	0.076469
LogisticRegression	0.610450	0.009227
RandomForestClassifier	0.774302	0.021827

Table 1. Criss Validation Scores (Mean Accuracy and Standard Deviation) of different Classification Algorithms

IV. Results and Discussion

To build the classifier, we used the implementation of scikit learn for the Random Forest Classifier. All of the default parameters were used, with the exception of the n_estimators parameter, which we set to 900 because, generally speaking, the more trees in the forest, the better the performance, but at the expense of a higher computational cost. A random forest is a meta-estimator that employs averaging to increase predictive accuracy and control overfitting after fitting numerous decision tree classifiers on various dataset subsamples. If bootstrap=True (the default), the sub-sample size is controlled by the max_samples parameter; otherwise, each tree is constructed using the entire dataset. The default parameters of the scikit learn's random forest classifier are: *n_estimators=100, *, criterion='gini', max_depth=None, min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features='sqrt', max_leaf_nodes=None, min_impurity_decrease=0.0, bootstrap=True, oob_score=False, n_jobs=None, random_state=None, verbose=0, warm_start=False, class_weight=None, ccp_alpha=0.0, max_samples=None.*

Aside from the simple accuracy score, we also used precision, recall, and F1-score measures to evaluate the performance of our model. The F1-score or F-score is defined as the harmonic mean of precision and recall.

$$F - score = \frac{precision * recall}{precision + recall}$$

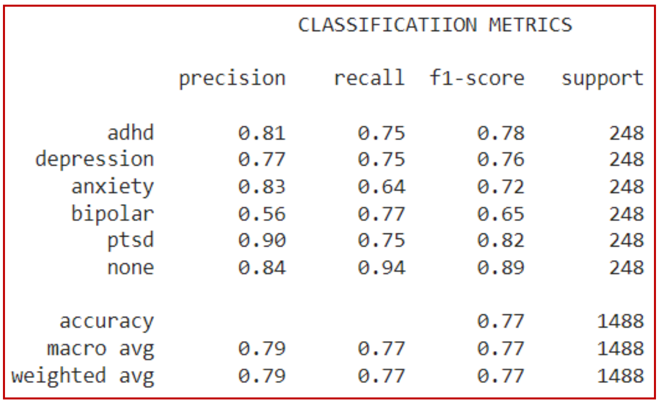
Precision and recall are defined as:

$$Precision = \frac{tp}{tp+fp}$$

$$Recall = \frac{tp}{tp+fn}$$

where tp, fp, and fn are true positives, false positives, and false negatives respectively.

The classification metrics can be seen in table 2 below. As we can observe on the table, the 'none' class has the highest F1 score with 0.89 followed by the PTSD class with 0.82. The bipolar class has the lowest F1-score with only 0.65. Overall, the classification performance is still pretty decent, it is not that bad or low.

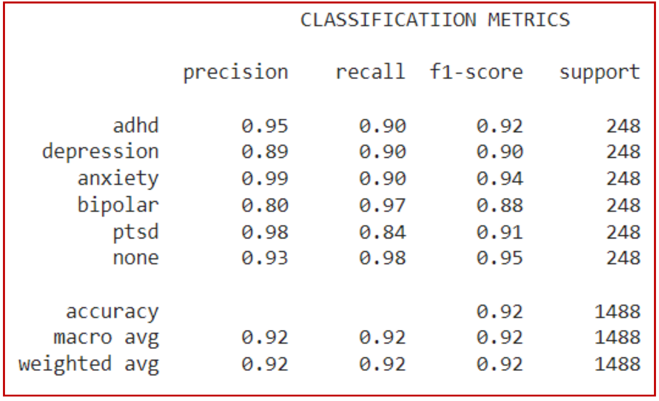


```
                     CLASSIFICATIION METRICS

              precision    recall  f1-score   support

        adhd       0.81      0.75      0.78       248
  depression       0.77      0.75      0.76       248
     anxiety       0.83      0.64      0.72       248
     bipolar       0.56      0.77      0.65       248
        ptsd       0.90      0.75      0.82       248
        none       0.84      0.94      0.89       248

    accuracy                           0.77      1488
   macro avg       0.79      0.77      0.77      1488
weighted avg       0.79      0.77      0.77      1488
```

Table 2. Classification Metrics

Training the model using Training-Validation Set

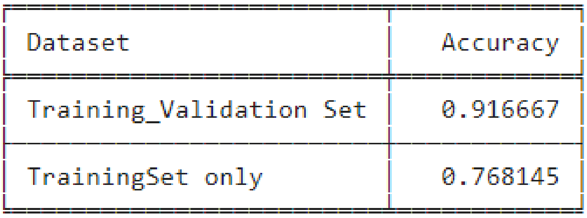
With the idea that a larger training set might improve the classifier model's performance, we also attempted to train the model using the concatenated dataset of the training and validation set with the same parameters. We also obtained the precision, recall, and F1 scores. The classification metrics utilizing the combined training and validation set are shown in Table 3. As we can see in the table, the result of the classification report using the combined dataset of the training and validation set is significantly better than using the training set alone to train the model. Additionally, we can observe that almost every class produced an F1 score greater than 0.9. Similar to the result of the classification metrics using the training set alone, the 'none' class also has the highest F1-score and the 'bipolar' class has the lowest.



```
                     CLASSIFICATIION METRICS

              precision    recall  f1-score   support

        adhd       0.95      0.90      0.92       248
  depression       0.89      0.90      0.90       248
     anxiety       0.99      0.90      0.94       248
     bipolar       0.80      0.97      0.88       248
        ptsd       0.98      0.84      0.91       248
        none       0.93      0.98      0.95       248

    accuracy                           0.92      1488
   macro avg       0.92      0.92      0.92      1488
weighted avg       0.92      0.92      0.92      1488
```

Table 3. Classification Metrics in using Training-Validation Set

Comparison of using Training Set and using Training-Validation Set



```
╒════════════════════════╤════════════╕
│ Dataset                │   Accuracy │
╞════════════════════════╪════════════╡
│ Training_Validation Set│   0.916667 │
├────────────────────────┼────────────┤
│ TrainingSet only       │   0.768145 │
╘════════════════════════╧════════════╛
```

Table 4. Accuracy Scores of Using training-validation set and training set only.

Table 4 shows the accuracy scores of the two datasets with the different sizes we experimented on. As we can see, the accuracy score of the model that is trained with the Training-Validation Set is far greater than that of the model that is trained using only the Training set. The training-validation set yielded an accuracy score of 92% while using only the training set yields an accuracy of 77%. From this experiment, we can say that with the data cleaning methods we applied, the features we extracted, and the model we used, we were able to build a good-performing classifier. We only need a larger dataset to improve its performance.

The confusion matrix for using only the training set and the confusion matrix for using the training-validation set is depicted in Figures 2 and 3 respectively. Both confusion matrix shows that the vast majority of the predictions end up on the diagonal (predicted label = actual label), where we want them to be. We can deduce that those two models, even with the model that uses only the training set that has a lower accuracy, perform very well.

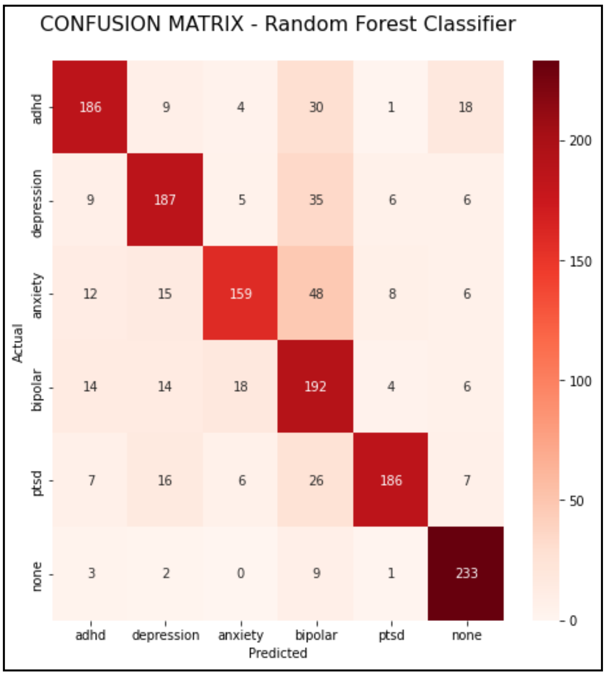


Figure 2. Confusion Matrix (training set)

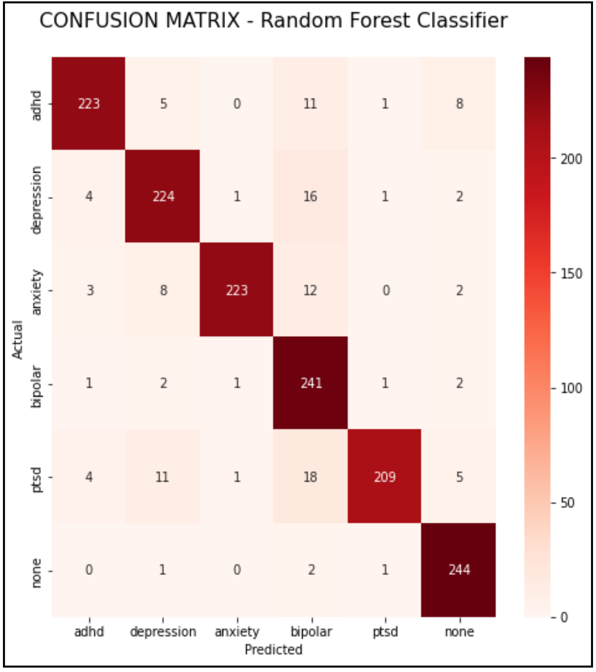
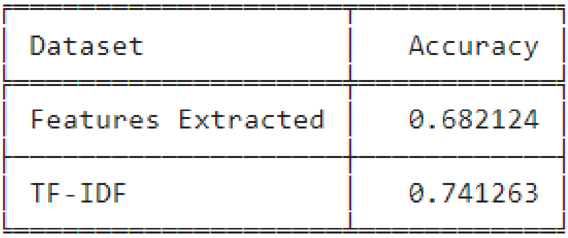


Figure 3. Confusion Matrix (training-validation set)

Comparison of TF-IDF and other Features

We also compared the accuracy of using only the TF-IDF to fit the model with the accuracy of the model obtained using the features we extracted without the TF-IDF, such as the total number of words, POS tag counts, text polarity, count of the frequency of each mental illness word, and frequency count of words that may signal or be related for a certain mental illness. Table 5 below shows a comparison of their accuracy scores. The accuracy score obtained when utilizing simply the TF-IDF is higher than when using the features, we extracted without the TF-IDF, as seen in the table. We can deduce that the feature we obtained from TF-IDF is superior to all the other features we extracted based on the fact that the accuracy score is almost 6% higher when using TF-IDF alone than when using all the other features combined. We believe that the reason why using only the features we extracted using TF-IDF is better, is because TF-IDF does not simply count the frequency of words in a document. It also creates a normalized count where each word count is divided by the number of documents this word appears in, unlike the other features we extracted that are almost just simple frequency counts of words and counts of Parts of Speech tags. TF-IDF also provides us with the importance of a term, which scales it based on how important it is across all documents after considering how important it is in a single document. From this observation, we can infer that the TF-IDF in the whole feature extraction greatly helps the final model to classify more accurately.



Dataset	Accuracy
Features Extracted	0.682124
TF-IDF	0.741263

Table 4. Comparison of Accuracy Scores of using the features extracted (without TF-IDF) and using only the TF-IDF.

References

[1] Jalal, N., Mehmood, A., Choi, G. S., &amp; Ashraf, I. (2022, March 31). A novel improved random forest for text classification using feature ranking and optimal number of trees. Journal of King Saud University - Computer and Information Sciences. Retrieved November 10, 2022, from https://www.sciencedirect.com/science/article/pii/S1319157822000969

[2] Orabi, A.H., Buddhitha, P., Orabi, M.H., Inkpen, D.: Deep learning for depression detection of twitter users. In: Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic. pp. 88–97 (2018)

[3] Abusaa, M., Diederich, J., Al-Ajmi, A., et al.: Machine learning, text classification and mental health. HIC 2004: Proceedings p. 102 (2004)

[4] Benton, A., Mitchell, M., Hovy, D.: Multi-task learning for mental health using social media text. arXiv preprint arXiv:1712.03538 (2017)

[5] Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K., Mitchell, M.: 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. pp. 31–39 (2015)

[6] Dinu, A., & Moldovan, A. C. (2021, September). Automatic detection and classification of mental illnesses from general social media texts. In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021) (pp. 358-366).

[7] Ameer, I., Arif, M., Sidorov, G., Adorno, H., & Gelbukh, A. (2022, July 3). Mental Illness Classification on social media texts using Deep Learning and Transfer Learning. Papers With Code.

[8] Breiman, L.: Random forests. Mach. Learn. 45, 5–32 (2001)

[9] Qi, Y., Bar-joseph, Z., Klein-seetharaman, J.: Evaluation of different biological data and computational classification methods for use in protein interaction prediction. Proteins 63, 490–500 (2006)

[10] Lebedev, A.V., Westman, E., Van Westen, G.J.P., et al.: Random Forest ensembles for detection and prediction of Alzheimer's disease with a good between-cohort robustness. NeuroImage Clin. 6, 115–125 (2014)

[11] Pflueger, M.O., Franke, I., Graf, M., Hachtel, H.: Predicting general criminal recidivism in mentally disordered offenders using a random forest approach. BMC Psychiatr. 15(1), 1 (2015)

[12] Rahman, M.M., Davis, D.N.: Machine learning-based missing value imputation method for clinical datasets. In: Yang, Gi-Chul, Ao, Sio-long, Gelman, Len (eds.) IAENG Transactions on Engineering Technologies, pp. 245–257. Springer, Dordrecht (2013)

[13] Cetin, M. S.: New approaches for data-mining and classification of mental disorder in brain imaging data, Dissertation (2015)

[14] Savitz, J.B., Rauch, S.L., Drevets, W.C.: Clinical application of brain imaging for thediagnosis of mood disorders: the current state of play. Mol. Psychiatr. 18(5), 528–539 (2013)

[15] vaderSentiment. (2020, May 22). PyPI. https://pypi.org/project/vaderSentiment/

[16] [Donvito, T. (2022, June 28). 13 Common Words and Phrases That May Signal Depression. The Healthy. https://www.thehealthy.com/mental-health/depression/words-phrases-sign-depression/

[17] Anxiety disorders - Symptoms and causes. (2018, May 4). Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/anxiety/symptoms-causes/syc-20350961

[18] Attention-deficit/hyperactivity disorder (ADHD) in children - Symptoms and causes. (2019, June 25). Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/adhd/symptoms-causes/syc-20350889

[19] Post-traumatic stress disorder (PTSD) - Symptoms and causes. (2018, July 6). Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/post-traumatic-stress-disorder/symptoms-causes/syc-20355967

[20] Bipolar disorder - Symptoms and causes. (2021, February 16). Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/bipolar-disorder/symptoms-causes/syc-20355955

[21] Murarka, A., Radhakrishnan, B., Ravichandran, S.: Classification of mental illnesses on social media using roberta. In: Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis. pp. 59–68 (2021)

[22] Cohan, A., Desmet, B., Yates, A., Soldaini, L., MacAvaney, S., Goharian, N. "SMHD: a Large-Scale Resource for Exploring Online Language Usage for Multiple Mental Health Conditions". Proceedings of the 27th International Conference on Computational Linguistics, Association for Computational Linguistics, 2018, pp. 1485–97. ACLWeb, https://www.aclweb.org/anthology/C18-1126.

[23] Jiang, Z., Levitan, S., Zomick J., Hirschberg, J., "Detection of Mental Health from Reddit via Deep Contextualized Representations". Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis, Association for Computational Linguistics, 2020, pp. 147–56, doi: 10.18653/v1/2020.louhi-1.16.

[24] Kim, J., Lee, J., Park, E., Han, J.: A deep learning model for detecting mental illness from user content on social media. Scientific reports 10(1), 1–6 (2020)

[25] Gkotsis, G., Oellrich, A., Velupillai, S., Liakata, M., Hubbard, T.J., Dobson, R.J., Dutta, R.: Characterisation of mental health conditions in social media using informed deep learning. Scientific reports 7(1), 1–11 (2017)

[26] Zirikly, A., Resnik, P., Uzuner, O., Hollingshead, K.: 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. pp. 24–33 (2019)

[27] C. Luo, Z. Wang, S. Wang, J. Zhang, J. Yu: Locating facial landmarks using probabilistic random forest. IEEE Signal Process. Lett., 22 (12) (2015), pp. 2324-2328

[28] A. Paul, D.P. Mukherjee: Mitosis detection for invasive breast cancer grading in histopathological images. IEEE Trans. Image Process., 24 (11) (2015), pp. 4041-4054

[29] Close T.M. Khoshgoftaar, M. Golawala, J. Van Hulse, An empirical study of learning from imbalanced data using random forest, in: 19th. IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007), Vol. 2, IEEE, 2007, pp. 310–317.

[30] X.-Y. Liu, J. Wu, Z.-H. Zhou: Exploratory undersampling for class-imbalance learning. IEEE Trans. Syst., Man, Cybern. Part B (Cybernetics), 39 (2) (2008), pp. 539-550

[31] T.G. Dietterich, Ensemble methods in machine learning, in: International workshop on multiple classifier systems, Springer, 2000, pp. 1–15.

[32] I. Ashraf, S. Hur, Y. Park: Magio: Magnetic field strength based indoor-outdoor detection with a commercial smartphone. Micromachines, 9 (10) (2018), p. 534

[33] A. Criminisi, J. Shotton: Decision forests for computer vision and medical image analysis. Springer Science & Business Media (2013)