# Choosing the accurate Dataset and Model

## 1. Understanding the Problem

### *1.1 Depression*

Depression is a widespread and severe global public health issue that cuts across demographic lines such as age, gender, and socioeconomic level. This complex mental health disorder is marked by a continuous sense of despair, hopelessness, and a significant lack of interest in previously enjoyable activities. Depression, if left untreated, can have far-reaching, negative implications, such as reduced everyday functioning, disrupted relationships, and, in its most severe manifestation, suicide.The magnitude of the problem is mind-boggling, with the World Health Organization estimating that nearly 280 million people worldwide suffer from depression. This statistic emphasizes the enormous impact of depression on both individuals and societies. However, a major source of concern is the large number of people suffering from depression who do not seek therapy or professional help According to the National Sample Survey 2017-'18, the average cost of hospitalization for psychiatric and neurological illnesses was Rs 26,843, with care in public hospitals costing Rs 7,235 and private hospitals costing Rs 41,239. According to a Mint column, the average cost of counselling or psychiatric consultation in any Indian metropolis is around Rs 1,500 per hour, while some mental health practitioners charge as much as Rs 2,000 to Rs 4,000 per hour.

Global provisions and services for identifying, supporting, and treating mental illness of this nature have been considered as insufficient (Detels, 2009). Although 87% of the world’s governments offer some primary care health services to tackle mental illness, 30% do not have programs, and 28% have no budget specifically identified for mental health (Detels, 2009). In fact, there is no reliable laboratory test for diagnosing most forms of mental illness; typically, the diagnosis is based on the patient’s selfreported experiences, behaviors reported by relatives or friends, and a mental status examination.

Thus we are in a huge need of AI Therapy. But according to Westmoreland Psychotherapy Associates. AI therapy may not be equipped to handle emergency situations effectively, and the absence of a human therapist can be a significant drawback during such times. An AI can’t provide timely intervention, reassurance, and emergency contacts in the same way a human can.However AI and Machine learning could be used for the diagnosis of such mental heath disorders, If trained with the right data.

## *1.2 Existing Research*

Researchers have shown great interest in using social media data to detect mental disorders. They have explored various platforms, including Twitter, Reddit, Sina Weibo, Facebook, and Instagram. Among the mental disorders and related symptoms, depression, suicidal ideation, schizophrenia, and eating disorders have been the focus of numerous studies (Chancellor & De Choudhury, 2020).

Machine learning techniques have been applied to analyze behavior patterns and linguistic styles on social media. For example, researchers found that increased expressions of sadness and the use of swear words are associated with depression (Rodriguez, Holleran, and Mehl, 2010). Depressed individuals tend to use more personal pronouns (e.g., "I") and verbs in the continuous, imperfective, or past tenses (Smirnova et al., 2018).

In one study, social engagement features, such as the normalized number of posts, proportion of reply posts, fraction of retweets, proportion of shared links, and fraction of question-centric posts, were used to detect depression on Twitter (De Choudhury, Counts, and Horvitz, 2013).

A Self-reported Mental Health Diagnoses (SMHD) dataset  was developed from Reddit posts (Cohan et al., 2018). They used a range of techniques, including Logistic Regression, Support Vector Machines (SVM), and XGBoost, as well as neural networks like Convolutional Neural Networks (CNN) and FastText to create a binary classifier for identifying individuals with depression.

Strube et al. employed a Hierarchical Attention Network (HAN) to identify depression in the SMHD dataset (Strube et al., 2019). In another study, Dinu et al. used three pretrained transformers (BERT, XLNET, and RoBERTa) to develop a binary classifier for detecting depression from Reddit posts at the post level, achieving high accuracy (Dinu et al., 2021).

In 2022, Vanessa et al. employed word embedding techniques like GloVe and Word2Vec to create domain-specific embeddings for Reddit posts in the SMHD dataset. They then integrated these embeddings into CNN and LSTM models to detect depression with good results (Vanessa et al., 2022).

Other researchers, such as Coppersmith et al. (2015) and Yates, Cohan, and Goharian (2017), used social media data to identify individuals with depression or PTSD by searching for self-reported diagnoses in tweets and Reddit posts, respectively.

However a lot of these models have several drawbacks even with a higher accuracy.

Pennebaker et al. (2007) utilized the Linguistic Inquiry Word Count (LIWC) to assess textual data for indicators of positive and negative affect, as well as various linguistic styles. LIWC is a well-established and validated tool for psychometric text analysis, consisting of a dictionary of 4,500 words or word stems categorized into groups such as adverbs, religion, and anxiety. Textual data is analyzed by measuring the extent to which these word categories are used. Additionally, they used the Affective Norms for English Words (ANEW) by Bradley and Lang (1999) to determine activation and dominance. Activation measures the intensity of emotions (e.g., "terrified" having higher activation than "scared"), while dominance represents the level of control over an emotion. For negative emotions, "anger" exhibits higher dominance than "fear," while for positive emotions, "optimism" has higher dominance than "relaxed." The study also examined linguistic features associated with depression, including a depression lexicon and antidepressant usage. The depression lexicon was constructed by mining a 10% sample of the "Mental Health" category on Yahoo! Answers and selecting the top 1% of terms with the highest mutual information and log likelihood ratio, based on the regex "depress*."

Kang, Yoon, and Kim (2016) developed a multimodal analyzer to predict user moods on Twitter. Their approach used text, emoticons, and images features in combination with a Support Vector Machine (SVM) classifier. They trained and evaluated the analyzers using manually collected datasets of tweets with positive and negative moods, as well as a well-known dataset of positive and negative reviews. Text and emoticon analyzers were trained on one dataset and validated on several others. The text analyzer included morphological analysis with a mood lexicon and a part-of-speech (POS) feature vector. The emoticon analyzer calculated polarity scores using an emoticon lexicon. An image dataset labeled as positive, neutral, and negative was used for training and validation of the image analyzer. The text, emoticon, and image models collectively achieved an F1 score of 0.8672 on one of the validation datasets.

Shen et al. (2017) collected a dataset to distinguish depressed Twitter users from non-depressed ones. They proposed a multimodal approach, employing social network, user profile, visual, emotional, topic-level, and domain-specific features. Social network features included tweet counts, followings, and followers. User profile features encompassed gender, age, relationships, and education levels. Visual features were derived from users' avatars. Emotional features consisted of sentiment analysis using LIWC and emotion words, a sentimental emoji library, and valence, arousal, and dominance features from ANEW. They also employed Latent Dirichlet Allocation (LDA) for topic modeling. Domain-specific features were word counts related to antidepressants and depression symptoms.

To identify Twitter users with Bipolar Disorder (BD) and Borderline Personality Disorder (BPD), Saravia, Chang, De Lorenzo, and Chen (2016) manually curated Twitter accounts associated with these disorders and used TF-IDF and pattern of life features with a Random Forest classifier. The TF-IDF features achieved a precision of 96% for both BD and BPD models, while pattern of life features yielded a precision of 91% for BD and 92% for BPD. Pattern of life features encompassed age, gender, polarity features, and social features. They used the Sentiment140 API to label tweets as positive, neutral, or negative, which were then transformed into affective features. Social features included daily posting frequency and the number of users mentioned multiple times.

### *1.3 Drawbacks of Previous Models*

To analyse the existing models I decided to go through several datasets and subreddits used in the making of the models and found several issues.

The existence of unnecessary talks among the gathered postings was the first big difficulty discovered.These discussions frequently diverted from the main issue of depression, making it difficult to identify significant content. I also discovered that therapists and mental health experts frequently contributed to discussions in mental health forums. This presented an issue because categorizing all postings from the forums as "depressed" is an oversimplification.

However even though a manual human annotator was not present in several researches, some researches had manually removed such data that was not related to depression.

I also identified several other notable issues. One predominant issue was the prevalence of comments with a limited number of characters. These posts responses often lacked the depth and substance required for meaningful analysis and posed a challenge in extracting substantial insights from them. However this can also be easily tackled by making the number of characters above a certain number.

However **the main issue** lies in determining the people that would be accessing the model. Depression detection models use the data classified as “not depressed” as a general conversation. For eg. Several papers use the subreddit “r/general”. Other papers remove words such as mental health from the data classified as “not depressed”. This does not create a depression detection model but instead creates a sentiment analysis model. This is because the prevalence of specific terms like "depressed" and "anxious" may lead models to rely heavily on a narrow set of words for categorization, which can reduce the accuracy of the classification. A person with anxiety can be classified as “depressed” which can be really dangerous.

Also, Individuals on social media platforms often self-report their emotional states, but these reports may not always be clear-cut or accurate. Users may misinterpret their own emotional states, leading to self-diagnoses that may not align with clinical assessments or other objective measures.

To analyse this further, We used a reddit dataset and a naïve bytes model to try to imitate the results of these research. Even though we obtained a 93% percent accuracy, The algorithm marked the statement “I am sad” as diagnosed with depression. This clearly elucidates the potential dangers of such model.

To tackle such issues, We created a comprehensive dataset for depression detection.

## 2. Dataset

The Dataset used was a combination of an existing open-source dataset and data scraped from reddit posts.

a. Existing Dataset **(Counsel-Chat.com)**

Counsel-chat is a platform where visitors can ask questions that experienced therapists will answer. They have graciously made their data (questions asked by the users) openly available to everyone. The data is in the form of a CSV, with each row having the question and the answer, as well as the name of the therapist answering and the question. There is also a row indicating the disorder/issue the post talks about. We removed all the columns except the question and the labelled  disorder, as the other columns were not necessary for our model. I marked all the labels stating depression as “1” and the others as “0”

Given below is a small extract from the dataset. There are over 2000 questions however a lot of them are repeated around 10 times. This will be handled in the preprocessing

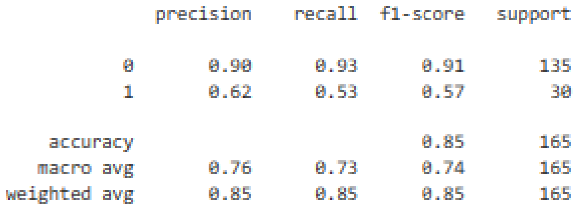
| Text | Depressed |
|---|---|
| I'm going through some things with my feelings and myself. I barely sleep and I do nothing but think about how I'm ... | 1 |
| I'm going through some things with my feelings and myself. I barely sleep and I do nothing but think about how I'm ... | 1 |
| I'm going through some things with my feelings and myself. I barely sleep and I do nothing but think about how I'm ... | 1 |
| I'm going through some things with my feelings and myself. I barely sleep and I do nothing but think about how I'm ... | 1 |

Even though this data was initially considered sufficient on its own and we had achieved an accuracy of 85%, The confusion matrix revealed the need for additional data

## b. Reddit Data

Our analysis of the counsel chat data revealed certain limitations that needed to be addressed. A prominent issue was the significant data imbalance between depressed and non-depressed individuals, with only about 150 rows pertaining to depression. When we initially applied the Naive Bayes model to this dataset, it yielded an accuracy of 84.8%. However, the low F1 score raised concerns.

Upon closer examination, we identified a stark disparity in the model's predictive performance. It excelled in predicting non-depressed cases but struggled when it came to identifying depression. The precision for depressed cases was approximately 0.62, while for non-depressed cases, it stood at 0.9. This discrepancy clearly indicated that the imbalanced data was affecting the model's effectiveness.



```
              precision    recall  f1-score   support

           0       0.90      0.93      0.91       135
           1       0.62      0.53      0.57        30

    accuracy                           0.85       165
   macro avg       0.76      0.73      0.74       165
weighted avg       0.85      0.85      0.85       165
```

Addressing this issue posed a challenge, as reducing data to achieve a balanced dataset wasn't feasible. To rectify this imbalance and enhance the model's ability to predict depression accurately, we made the strategic decision to augment the dataset with additional data.

To achieve this, we curated comments from Reddit posts specifically titled "What does depression feel like?" Manually filtering out irrelevant information, as well as common words like "feel" and "It," we ensured that the added data was pertinent to the context. This augmentation proved to be a crucial step in improving the model's performance and addressing the imbalanced data issue.

## *2.1 Preprocessing*

The preprocessing approach was conducted on the column “Text” (all textual contents including posts, comments and titles). The preprocessing was done in 2 ways. First was manual preprocessing done manually in excel. The second was preprocessing done via python and libraries such as pandas and matplotlib.

### a) Manual Preprocessing

1. Remove URLs. Many users tend to include URLs in their online posts and comments; although some words in these URLs might be helpful to the classification task, it is difficult to distinguish the meaningful ones from random strings, so the complete URLs were removed

2. Convert text to lowercase characters. The lowercase text can reduce variations of words caused by capitalization.

3. Replace four or more consecutive repeating characters in a word with one character. Online posts and comments contain plenty of words using repeating characters to express strong feelings (e.g., “noooooo”). These words need to be normalized for an effective model

4. Remove words referring to mental health disorders such as Depression, Anxiety, PTSD, MDD etc. These words need to be removed for reducing the chances of self-diagnosis and our model should not rely on these words and instead rely on the words that talk about the experience of the person suffering from the issues

5. Remove words talking about general emotions such as “happy” or “sad”. However, not removing words that talk about a specific emotion such as “angry”, “numb” etc. This is done so that the model does not classify general sentiments as a mental heatch disorder. However specific emotions are kept as they serve as crucial indicators for identifying potential mental health concerns.

6. Removing therapist inputs and unnecessary conversations.

7. Remove punctuation marks. As a lot of our data was questions, Removing punctuation marks was a must for an effective model

**b) Preprocessing in Python**

*1. Loading the Dataframe.*

We start by loading the data into a Pandas DataFrame. This is a crucial step for initiating the project.

2.*Removing Columns that are no longer needed*

By removing extraneous columns, we ensure that the analysis or model is focused on the most significant and useful attributes, resulting in more accurate and interpretable results. We removed columns such as “therapist-name”, qualification, response and the links of the questions

*3.Handling Values That Are Duplicated and Missing*

Correctly handling missing data helps in the prevention of bias in the analysis. Ignoring missing values or incorrectly treating them can result in inaccurate conclusions

## *2.2 Exploratory Data Analysis*

Summary Statistics for Text Data: To begin, I computed summary statistics for three main properties in the dataset: 'char' (character count), 'words' (word count), and 'sentences' (sentence count). These statistics are useful for acquiring a high-level knowledge of text data distribution and features. The percentiles (25%, 50%, and 75% of the total) show how these traits are spread across the dataset.

Statistics for Depressed vs. Non-Depressed People:

To acquire a better understanding of the data, I divided it into two unique groups based on the 'Depressed' column. I first calculated summary statistics for the subset where 'Depressed' equals 0 and then for the subset where 'Depressed' equals 1.

This division enables me to compare and contrast the character, word, and sentence counts of persons who are and are not depressed.

Data Filtering: Data filtering is an important part of data preprocessing. I filtered the dataset to maintain consistency and guarantee that the study focuses on a more consistent range of text lengths. I specifically removed rows with a character count of less than 1500 in the 'Text' column. This phase is critical for removing any outliers and ensuring that the analysis and subsequent machine learning models run on data with comparable text length ranges.

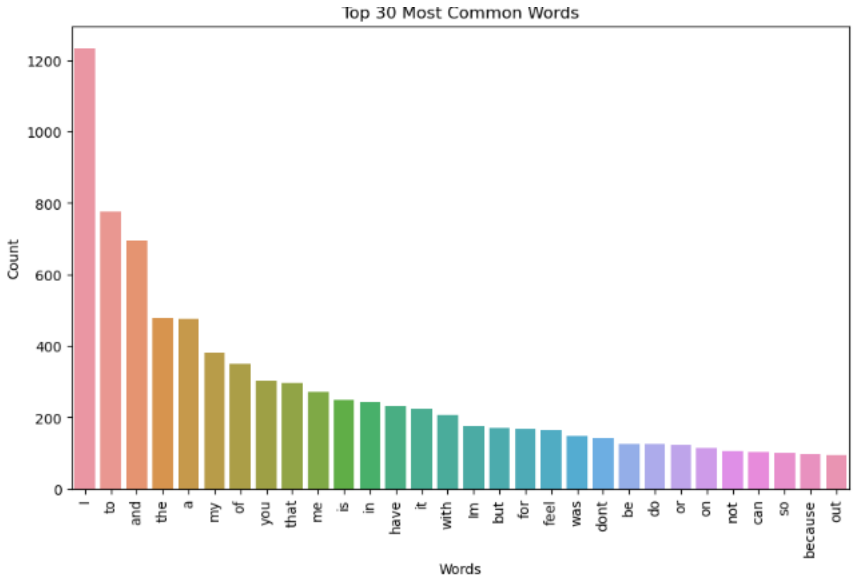
Word Frequency Analysis: I conducted a word frequency analysis on the text data of individuals classified as 'Depressed' (where 'Depressed' is equal to 1).

Here's a detailed breakdown of the process:

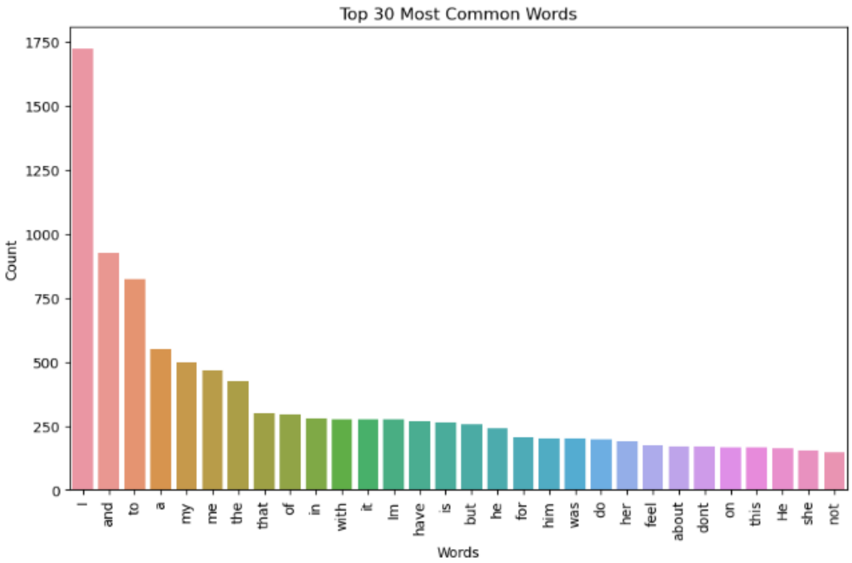
1.I initialized an empty list named 'words' to capture individual words from the text data. I iterated through the text data of depressed individuals, breaking each text into individual words and appending these words to the 'words' list.

2.Utilizing the Python collections.Counter module, I tallied the frequency of each unique word in the 'words' list.

3.I created a DataFrame to organize the top 30 most common words and their corresponding frequencies. To provide a visual representation of the findings, I generated a barplot using the 'seaborn' library. This barplot displays the top 30 most frequently occurring words in the text data for Depressed and Non Depressed Text data.



The above graph shows the top 30 Most common words in Non-Depressed Individuals



This graph shows the most common words for Depressed Individuals. We can clearly see, as stated by other studies, the excessive use of personal pronouns in Depressed individuals. For eg. The frequency of the word “I” is about 1750 in depressed individuals and 1200 in non-depressed individuals

## 3. Data Modelling

In the process of creating our text classification model for detecting signs of depression, we followed a systematic and comprehensive approach. Our dataset, containing text data and corresponding labels indicating whether an individual is depressed or not, served as the foundation for our model development.

First, we extracted the textual data (X) and the corresponding labels (y) from our dataset. We then employed the 'train_test_split' function from the 'sklearn.model_selection' module to divide our dataset into training and testing sets. This step was essential to assess the model's performance. The training set (X_train and y_train) comprised 80% of the data, while the remaining 20% was allocated to the testing set (X_test and y_test). The 'random_state' parameter was set to ensure reproducibility of our results.

Upon inspecting the dimensions of our training and testing sets, it was evident that our training set consisted of 624 samples, while the testing set contained 157 samples. These figures indicate the data split we had achieved, laying the groundwork for training and evaluating our model.

To prepare our textual data for the machine learning model, we utilized the 'CountVectorizer' from the 'sklearn.feature_extraction.text' module. The 'CountVectorizer' is a critical tool for text preprocessing, as it converts the raw text into a numerical format that machine learning algorithms can work with. It transforms the text data into a sparse matrix, representing the frequency of words or tokens within the text. The result is an array that shows the occurrence of each word or token in the dataset.

This initial transformation of the text data was a crucial step in our model creation process, as it allowed us to quantify the textual information for the machine learning algorithm. Once we had these numerical representations of the text, we were ready to proceed with model training.

In the Model Training we decided to test several models such as Naïve Bytes, Random foret, XG Boost, MLP Classifier etc and noted the results.

## 4. Results

In this experimental study, we assessed the performance of various machine learning models in predicting depression based on textual data. The findings revealed that the Multinomial Naive Bayes model stood out with the highest testing accuracy of nearly 79% and an F1 score of 0.7402. This model achieved a balanced combination of precision and recall, making it a strong contender for depression prediction in this context. The first table shows the results of the models before lemmatization and the second shows the results after lemmatization

After lemmatizing the text data, the Multinomial Naive Bayes model continues to perform the best with a testing accuracy of 78.98% and an F1 score of 0.7402. XGBoost also performs well, with a testing accuracy of 73.89% and an F1 score of 0.6917. The MLP Classifier achieved a testing accuracy of 71.34% and an F1 score of 0.6980. These results indicate that lemmatization has had a positive impact on the models' performance, particularly for the Multinomial Naive Bayes model.

| Model | Testing Accuracy | F1 Score | Precision | Recall |
|---|---|---|---|---|
| Multinomial Naive Bayes | 0.7898 | 0.7402 | 0.790 / 0.780 | 0.860 / 0.700 |
| Random Forest | 0.7771 | 0.7154 | 0.780 / 0.790 | 0.870 / 0.660 |
| Logistic Regression | 0.7389 | 0.6963 | 0.780 / 0.690 | 0.770 / 0.700 |
| K-Nearest Neighbors (k=2-20) | 0.7389 | 0.6963 | - | - |
| XGBoost | 0.7389 | 0.6917 | - | - |
| Support Vector Machine | 0.6051 | 0.225 | - | - |
| MLP Classifier | 0.7261 | 0.6667 | 0.750 / 0.690 | 0.790 / 0.640 |

| Model | Testing Accuracy | F1 Score |
|---|---|---|

| | | |
|---|---|---|
| Multinomial Naive Bayes | 0.7898 | 0.7402 |
| XGBoost | 0.7389 | 0.6917 |
| MLP Classifier | 0.7134 | 0.6980 |

## 5. References and citations

Arvind, Banavaram Anniappan, et al. "Prevalence and Socioeconomic Impact of Depressive Disorders in India: Multisite Population-Based Cross-Sectional Study." BMJ Open, vol. 9, no. 6, June 2019, p. e027250. PubMed Central, https://doi.org/10.1136/bmjopen-2018-027250.

Gwynn, R. Charon, et al. "Prevalence, Diagnosis, and Treatment of Depression and Generalized Anxiety Disorder in a Diverse Urban Community." Psychiatric Services, vol. 59, no. 6, June 2008, pp. 641–47. ps.psychiatryonline.org (Atypon), https://doi.org/10.1176/ps.2008.59.6.641.

Hamblion, Esther, et al. "Global Public Health Intelligence: World Health Organization Operational Practices." PLOS Global Public Health, edited by Saskia Popescu, vol. 3, no. 9, Sept. 2023, p. e0002359. DOI.org (Crossref), https://doi.org/10.1371/journal.pgph.0002359.

Rude, Stephanie, et al. “Language Use of Depressed and Depression-Vulnerable College Students.” Cognition & Emotion, vol. 18, no. 8, Dec. 2004, pp. 1121–33. DOI.org (Crossref), https://doi.org/10.1080/02699930441000030.

Saravia, Elvis, et al. “MIDAS: Mental Illness Detection and Analysis via Social Media.” 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2016, San Francisco, CA, USA, August 18-21, 2016, edited by Ravi Kumar et al., IEEE Computer Society, 2016, pp. 1418–21. DBLP Computer Science Bibliography, https://doi.org/10.1109/ASONAM.2016.7752434.

Therapy, W. P. A. “A Chatbot for AI Therapy? The Drawbacks of ChatGPT for Mental Health.” WPA, 9 July 2023, https://www.wpatherapy.com/post/a-chatbot-for-ai-therapy-the-drawbacks-of-chatgpt-for-mental-health.

Buddhitha, Prasadith, and Diana Inkpen. “Multi-Task Learning to Detect Suicide Ideation and Mental Disorders among Social Media Users.” Frontiers in Research Metrics and Analytics, vol. 8, Apr. 2023, p. 1152535. PubMed Central, https://doi.org/10.3389/frma.2023.1152535.

Chancellor, Stevie, and Munmun De Choudhury. “Methods in Predictive Techniques for Mental Health Status on Social Media: A Critical Review.” Npj Digital Medicine, vol. 3, no. 1, Mar. 2020, pp. 1–11. www.nature.com, https://doi.org/10.1038/s41746-020-0233-7.

Dinu, Anca, and Andreea-Codrina Moldovan. “Automatic Detection and Classification of Mental Illnesses from General Social Media Texts.” Proceedings of the International

Conference on Recent Advances in Natural Language Processing (RANLP 2021), edited by Ruslan Mitkov and Galia Angelova, INCOMA Ltd., 2021, pp. 358–66. ACLWeb, https://aclanthology.org/2021.ranlp-1.41.

Keumhee Kang, et al. “Identifying Depressive Users in Twitter Using Multimodal Analysis.” 2016 International Conference on Big Data and Smart Computing (BigComp), Jan. 2016, pp. 231–38. Semantic Scholar, https://doi.org/10.1109/BIGCOMP.2016.7425918.

Rodriguez, Aubrey J., et al. “Reading between the Lines: The Lay Assessment of Subclinical Depression from Written Self-Descriptions.” Journal of Personality, vol. 78, no. 2, 2010, pp. 575–98. APA PsycNet, https://doi.org/10.1111/j.1467-6494.2010.00627.x.

Rude, Stephanie, et al. “Language Use of Depressed and Depression-Vulnerable College Students.” Cognition & Emotion, vol. 18, no. 8, Dec. 2004, pp. 1121–33. DOI.org (Crossref), https://doi.org/10.1080/02699930441000030.

Shen, Guangyao, et al. Depression Detection via Harvesting Social Media: A Multimodal Dictionary Learning Solution. 2017, pp. 3838–44. www.ijcai.org, https://www.ijcai.org/proceedings/2017/536.

Yates, Andrew, et al. Depression and Self-Harm Risk Assessment in Online Forums. arXiv, 6 Sept. 2017. arXiv.org, https://doi.org/10.48550/arXiv.1709.01848.

Counsel Chat Website: counselchat.com

Counsel Chat Dataset: https://github.com/nbertagnolli/counsel-chat

Kaggle Dataset https://www.kaggle.com/datasets/infamouscoder/depression-reddit-cleaned/code

Reddit Question https://www.reddit.com/r/AskReddit/comments/838ywu/what_does_depression_feel_like/