Winter 2020

Social Media Based Algorithmic Clinical Decision Support Learning from Behavioral Predispositions

Radhika V. Medury

SOCIAL-MEDIA BASED ALGORITHMIC CLINICAL DECISION SUPPORT LEARNING FROM BEHAVIORAL PREDISPOSITIONS

by

Radhika V. Medury, M.S.

A Dissertation Presented in Partial Fulfillment
of the Requirements of the Degree
Doctor of Philosophy

COLLEGE OF ENGINEERING AND SCIENCE
LOUISIANA TECH UNIVERSITY

March 2020

LOUISIANA TECH UNIVERSITY

THE GRADUATE SCHOOL

_____
Date

We hereby recommend that the dissertation prepared under our supervision by **Radhika V. Medury,** M.S.

entitled **Social Media Based Algorithmic Clinical Decision Support Learning from Behavioral Predispositions**

be accepted in partial fulfillment of the requirements for the Degree of

Doctor of Philosophy in Computational Analysis and Modeling

_____
Supervisor of Dissertation Research

_____
Head of Department

_____
Department

Recommendation concurred in:

_____

_____ Advisory Committee

_____

_____

Approved: **Approve**d:

_____ _____
Director of Graduate Studies Dean of the Graduate School

_____
Dean of the College

GS Form 13
(8/10)

ABSTRACT

Behavioral disorders are disabilities characterized by an individual's mood, thinking, and social interactions. The commonality of behavioral disorders amongst the United States population has increased in the last few years, with an estimated 50% of all Americans diagnosed with a behavioral disorder at some point in their lifetime. Attention-Deficit/Hyperactivity Disorder is one such behavioral disorder that is a severe public health concern because of its high prevalence, incurable nature, significant impact on domestic life, and peer relationships. Symptomatically, in theory, ADHD is characterized by inattention, hyperactivity, and impulsivity. Access to providers who can offer diagnosis and treat the disorder varies by location.

The ever-increasing use of social media can be effectively employed in the diagnosis and treatment of the disorder. Study of behavior and in extension, the study of individuals with behavioral disorders is made easier through the uninhibited setting in which posts are created on social media platforms.

Outside the United States, diagnosis rates of the disorder are low, as it is mainly considered to be an American disorder. This impression was reinforced by the perception that the disorder is caused by social and cultural factors common to American society. However, in reality, the disorder can as quickly affect people of different races and cultures worldwide, but recognition of the disorder in the medical community has been slow. This may be due to its adverse impact on an individual, their families, and society.

This dissertation focuses on providing clinicians with a clinical decision support system to overcome the societal stigma associated with the disorder and to ensure the accurate and efficient diagnosis of individuals with the disorder. The results provided in this dissertation assist in the diagnosis of individuals with Attention Deficit Hyperactivity Disorder. Data for individuals with the disorder is collected through posts of self-reported diagnoses on Twitter using the Twitter API. Previous research has proved that there are differences in behavior before and after the diagnosis of the disorder. To capitalize on this, symptomatic differences of the disease before and after diagnosis are discovered and evaluated. The symptoms of the disorder, namely, inattention, hyperactivity, and impulsivity, are quantified using measures of sentiment and semantics. A separate group of users without the disorder, the control group, are collected for validation. The analysis poses a three-class classification problem, with the classes being pre-diagnosed, post-diagnosed, and control groups. Decision trees are used to force all possible outcomes in the semantic and sentiment differences in the three classes of users to create a clear delineation. Behavioral disorders diagnosed by a clinician are based on identifying whether a patient deviates from an identified normal. This is evaluated by answering a set list of questions that quantify behavior. To achieve the same without manual intervention, ease in interpretability - decision trees are chosen. Classification using a decision tree is on a tweet-level and a user-level. Four cases are used both analyses: pre-diagnosed vs. post-diagnosed group, pre-diagnosed vs. control group, post-diagnosed vs. control group, and pre-diagnosed vs. post-diagnosed vs. control group.

The analysis on a user-level provides a higher degree of accuracy, with 93% accuracy for the case post-diagnosed vs. control group. The accuracy of the cases identifies

the number of people who can be correctly classified into their respective groups. Low accuracy for the tweet-level results fortifies the opinion that the sparsity of information in tweet level analysis is a disadvantage. This is overcome by analyzing on a user level. The accuracy of the classifier can be further improved upon by the addition of features such as age and gender. The addition of these features may also be useful in predicting time to remission and peak of the disorder in future studies.

APPROVAL FOR SCHOLARLY DISSEMINATION

The author grants to the Prescott Memorial Library of Louisiana Tech University the right to reproduce, by appropriate methods, upon request, any or all portions of this Dissertation. It is understood that "proper request" consists of the agreement, on the part of the requesting party, that said reproduction is for his personal use and that subsequent reproduction will not occur without written approval of the author of this Dissertation. Further, any portions of the Dissertation used in books, papers, and other works must be appropriately referenced to this Dissertation.

Finally, the author of this Dissertation reserves the right to publish freely, in the literature, at any time, any or all portions of this Dissertation.

Author _____

Date _____

DEDICATION

This dissertation is dedicated to my late grandfather, Dr. M. B. Rao. I hope this humble attempt makes you proud.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisor Dr. Sumeet Dua for his guidance and patience. To Normal John Mapes Jr., Dr. Pradeep Chowriappa and Anna White – thank you for making this dissertation possible. I would also like to thank the rest of my committee members – Dr. Weizhong Dai, Dr. Jean Gourd, and Dr. Jinko Kanno.

To my husband and my daughter, William and Aanya, thank you for patiently putting up with me while I spent all my time and effort on this dissertation. To my father, Dr. Yajulu Medury; my aunt, Dr. Manjula Guru; my uncle, Dr. Prasad Medury; my cousin, Dr. Devi Akella – you all are an inspiration. A big thanks to my brother, Ananta, for constantly badgering me to finish my doctorate. Finally, to my mother, my in-laws, my siblings, my grandmother, my aunts and my uncles – much love and gratitude.

CHAPTER 1

INTRODUCTION

Social media are websites or applications that enable users to create/share content or to participate in social networks. In the last few years, social media platforms such as Facebook, Twitter, and Instagram have been widely used, providing researchers with repositories of public data to be analyzed. The available public data may be in the form of messages, images or videos, and can provide real-time insight into public sentiment, general day-to-day activities, or events across the country or the world. For example, Cheong and Cheong at RMIT university identified vital players in existing online networks on Twitter during the 2010-2011 floods in Australia and generated new online networks to disseminate critical information (Cheong & Cheong, 2011).

Detection and dissemination of information related to public health have relied on social media as of late. The reason behind this is that the detection of public health threats through disease surveillance strategies using data transmitted from healthcare facilities, physicians has its limitations. Such data collection strategies take time, and context information on individual cases is often lost in transmission. To overcome such limitations, social media data has been exploited to detect, track, and disseminate health outbreaks. For example, Paul and Dredze analyzed public tweets and discovered mentions of various ailments such as allergies, obesity, and insomnia (Paul & Dredze, 2011). The illnesses were analyzed by geographic region, measuring risk factors, symptoms, and medication usage.

This dissertation strives to provide clinicians with a clinical decision support system to overcome the societal stigma associated with Attention Deficit Hyperactivity Disorder to ensure the accurate and efficient diagnosis of individuals with the same. This is achieved by identifying symptomatic differences in the disorder, before and after diagnosis by:

1. Establishing correlations in language and emotion by geographical prevalence of the disorder.
2. Establishing measures of disorder to quantify human behavior in terms of sentiment and semantics.
3. Developing a social media based clinical decision support system to aid in the accurate and efficient diagnosis of the disorder using supervised learning.

1.1 Data Mining

Data mining is defined as the process of finding hidden patterns from abundant data sources (Han, et al., 2000). The data can be databases, data warehouses, streaming data, or other information repositories. A term synonymous with data mining is Knowledge Discovery from Data (KDD). Figure 1-1 shows the iterative sequence of the steps involved in the KDD process.

1. **Data Selection**: This step involves retrieving data from existing data sources. The retrieved data may be further preprocessed to select a subset of attributes or features that may be relevant to the task at hand.
2. **Data Preprocessing**: The step involves the removal of noisy data such as errors, outliers, and inconsistent data. It may also include the integration of multiple data sources to enhance the efficiency of data mining further.

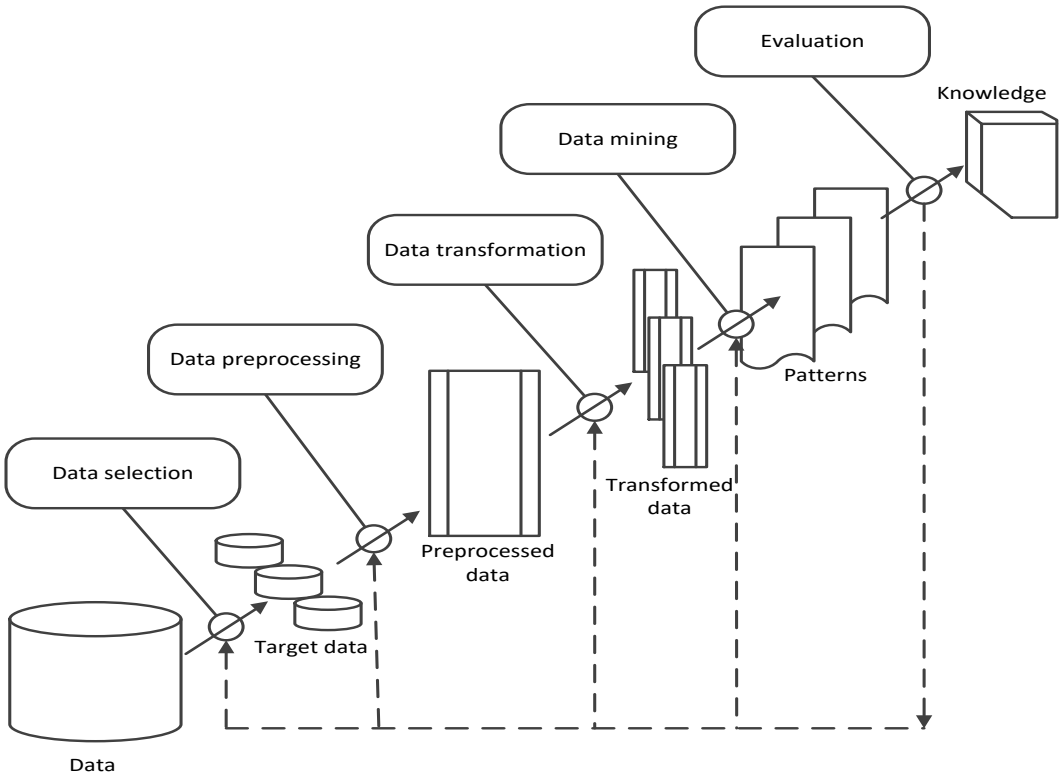

Figure 1-1: Knowledge discovery of data.

3. **Data Transformation**: This step involves the transformation and consolidation of data into forms that are deemed appropriate for the data mining task. Sub-tasks may include data normalization or discretization, feature construction, and data smoothing.
4. **Data Mining**: This step involves the application of data modeling techniques to extract hidden patterns from the target data in step 1.
5. **Evaluation**: The steps involve analyzing the extracted patterns to represent knowledge obtained from the target data successfully. Knowledge is then presented using visualization techniques to users of the system.

The steps shown above are collectively referred to as data mining in the industry. Social media data is per the five V's of big data:

1. **Volume**: Volume refers to the size of the data sets that need to be analyzed and processed.
2. **Variety**: Social media data is structured as well as unstructured.
3. **Velocity**: Velocity refers to the frequency of incoming data.
4. **Veracity**: Veracity refers to the trustworthiness of the data.
5. **Value**: Value refers to whether the collected data can provide any hidden insights.

1.2 Linear Regression

Linear regression is a smoothing technique that involves finding the best line to fit two attributes/variables so that one of the attributes can be used to predict the other (Han, et al., 2000). For example, a random variable, *y*, called a response variable, can be modeled as a linear function of another random variable, *x*, called a predictor variable, as follows:

$$y = wx + b \qquad \textbf{Eq 1.1}$$

where *w* and *b* are the regression coefficients. In the above equation 1.1, it is assumed that the variance of *y* is constant. The regression coefficient, *b*, is used to specify the slope of the y-intercept, and the regression coefficient, *w*, is used to specify the slope of the line. The two coefficients can be solved by using the method of least squares. The method of least squares minimizes the error between the estimate of the line and the actual line separating the data.

Linear regression can be used on sparse data sets, although its applicability may be limited. It handles skewed datasets exceptionally well, but when applied to high-dimensional data, it is computationally intensive.

1.3 Unsupervised Learning

Unsupervised learning, or clustering, is the term used when the learning process is unsupervised because the class labels are undefined (Han, et al., 2000). Clustering methods can be compared using the following aspects:

1. **Partitioning Criteria**: Objects may be partitioned into clusters such that either no hierarchy exists amongst the clusters; or into clusters at different semantic levels. Clusters with a hierarchy among them are used in text mining. For example, hierarchy is essential when performing topic detection on a corpus of documents.
2. **Separation of Clusters**: Objects may be partitioned into mutually exclusive clusters, or data points may belong to multiple clusters. The latter is used when clustering documents according to their topics, multiple topics may define a document.
3. **Similarity Measure**: Similarity between clusters can be calculated based on the distance between them; or maybe defined based on connectivity, density, contiguity. Both similarity measures play a significant role in the design of the clustering methods: distance-based methods use optimization techniques, and density/continuity-based methods can find clusters with no particular shape.
4. **Clustering Space**: Clustering methods that look for clusters in the entire given space are useful for low-dimensionality datasets. However, with high-dimensional data, such clustering methods lead to irrelevant data attributes making similarity measures unreliable. Therefore, it is advantageous to search for clusters in sub-spaces of the dataset.

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) relies on the density-based notion of clustering to find clusters of arbitrary shape in spatial databases with noise. The basic idea of this method is to group together data points in high-density areas and to mark data points in low-density regions as outliers. The density at a local point p is defined by two parameters: radius for the neighborhood of p, ϵ, and all the points from p within a radius ϵ, ϵ-neighborhood.

$$N_{\epsilon}(p) \coloneqq \{q \; in \; dataset \; D | dist(p,q) \leq \epsilon \qquad \textbf{Eq 1.2}$$

where q is a data point within radius ϵ of point p. In the neighborhood N(p), the minimum number of points is *MinPts*. If a ϵ-neighborhood contains at least *MinPts*, then the area is a high-density area (Ester, et al., 2003).

If point p is a core point and the point q is in the ϵ-neighborhood of point p, then the point q is directly density-reachable from the point p. If points, p, and q, are commonly density-reachable from a point o, then they are density-connected (Ester, et al., 2003).

The DBSCAN algorithm does not work well with areas of varying densities.

1.4 Supervised Learning

Supervised learning, or classification, is a term used for learning processes where the class labels are known (Witten, et al., 2016). Classifiers predict categorical class labels. Most classification algorithms are memory resident (a small data size). Typically, data classification is a two-step process: a learning process and a classification process. The learning step is where the classification model is constructed, and the classification process is where the model is used to predict the class labels for a given dataset.

In the learning step, the algorithm builds a classifier by learning from a training set made up of database tuples and associated class labels. The accuracy of a classifier, the

predictive accuracy of the said classifier is estimated. If the training set is used to predict the accuracy, the classifier tends to overfit the data. Therefore, a test set (independent of the training set) is used to predict the accuracy of the classifier. The accuracy of a classifier is then measured by the percentage of tuples in the test set that has correctly classified by the classification algorithm.

A decision tree is a structure that resembles a flow chart, where each non-leaf node represents an attribute, a branch represents an outcome, and each leaf node represents a class label. The node at the top is called the root node.

Decision trees are used for classification. If a tuple X is given, with an unknown class label, the attribute values for the given tuple are tested against a decision tree. A path from the root to the leaf is traced, where the leaf node holds the prediction for X. An advantage of decision trees is that they can be converted into classification rules easily. Other advantages of decision trees are they do not require any domain knowledge, can handle multidimensional data, and the learning/classification steps are fast and straightforward.

1.5 Conclusion

The chapter explains data mining, supervised learning, and unsupervised learning, touching upon the methodologies used in the chapters. The difference between supervised learning and unsupervised learning is majorly in the class labels being known/unknown. In this dissertation, the DBSCAN algorithm is the algorithm implemented for unsupervised learning, and a Decision tree is an algorithm implemented for supervised learning. DBSCAN has been implemented in Chapter 2, Neural Networks in Chapter 3, and Decision Trees in Chapter 4.

CHAPTER 2

CORRELATIONS IN LANGUAGE AND EMOTION FOR GEOGRAPHIC ADHD PREVALENCE

Behavioral disorders are an emotional disability that affects an individual's mood, thinking, and social interactions (CDC - Mental Health, 2019). The commonality of behavioral disorders amongst the United States population has increased in the last few years, with an estimated 50% of all Americans diagnosed with a behavioral disorder at some point in their lifetime (CDC - Data and Publications, 2018). Attention-Deficit/Hyperactivity Disorder (ADHD) is one such behavioral disorder that is a severe public health concern because of its high prevalence, incurable nature, significant impact on domestic life, and peer relationships (Hulkower, 2016).

Symptomatically, in theory, ADHD is characterized by inattention, hyperactivity, and impulsivity. In practicality, disorders such as Anxiety Disorders, Depression, and Bipolar Disorder may be biologically, physiologically, and emotionally like ADHD or in addition to ADHD. The severity of a person's behavioral disorder(s) determines whether he/she may in further risk of developing other diseases; for example, a person diagnosed with ADHD and Anxiety Disorder may have a high risk of developing diabetes. Correlations between language and emotion have previously proven to be effective in identifying and addressing factors that may significantly reduce the risk of developing such diseases (Eichstaedt, et al., 2015).

This chapter explores the use of social media data, mainly Twitter, to find correlations between language use of people diagnosed with ADHD and emotions using regression and cross-validation. The chapter is divided into four main sections, namely, related works, methodology, results, and conclusion.

2.1 Related Works

The use of social media to assist in learning about the personal, psychological, and behavioral aspects of communities has been explored in the past. Social media contains rich information in text, traits, preferences, and opinions (Volkova, et al., 2015). Durme (2012) showed that gender could be accurately predicted from Twitter language usage; Zamal, et al. (2012) predicted age; and Volkova, et al. (2014) predicted political views. Social media has also been used to understand emotional and mood changes over time in communities, for example, changes in emotional reactions over happy or sad events.

Sentiment and Semantic analysis have played a significant part in quantifying measures to identify and understand the correlates of behavioral disorders. De Choudhury, et al. (2013b) was one of the first to explore the use of Twitter to characterize Depression into quantifiable behavioral measures. Google researched trends in influenza by using search queries, successfully providing information on the onset of the ailment (Ginsberg, et al., 2009). Similarly, Twitter has been in other studies to track Lyme disease (Seifter, et al., 2010), H1N1 influenza (Chew & Eysenbach, 2010).

Cloninger, et al. (2006) explored the personality traits of individuals to predict future episodes of depression. Rude, et al. (2003), and Robinson and Alloy (2003) concluded that negative processing biases could predict depression by resolving ambiguous language. Moreno, et al. (2011) proved that Facebook status updates could reveal

symptoms of depressive episodes. Rude, et al. (2004) used LIWC to analyze written text to establish cues about neurotic tendencies and psychiatric disorders. De Choudhury, et al. (2013a) built a statistical model to examine behavioral changes in postnatal mothers by analyzing linguistic and emotional characteristics.

2.2 Methodology

The data used for this step has been taken from two sources: a CDC survey (Data and Statistics about ADHD, 2019) and research published by the NIH, "*Psychological language on Twitter predicts county-level heart disease mortality*" (Eichstaedt, et al., 2015). The latter explores language patterns on Twitter to identify community-level psychological correlates of age-adjusted mortality from Atherosclerotic Heart Disease (AHD) (Eichstaedt, et al., 2015). The former is an estimate of the state-wise prevalence of ADHD of youth aged 4-17 in the year 2011. The CDC data has two sections: diagnosis data and treatment data. The two sections are further subdivided into ever diagnosed, currently diagnosed, medicated, and diagnosed and medicated.

The data acquired from the NIH research is a comprehensive county-wise list of the relative frequency of language variables. The language variables are quantified as eight emotions: anger, engagement, disengagement, negative emotion, positive emotion, negative relationship, positive relationship. Additionally, it also provides a county-level measure of socioeconomic status (income and education), demographics (percentage of Black, Hispanic, married, and female residents) and health variables (incidence of diabetes, obesity, smoking, and hypertension).

2.2.1 Algorithms, Definitions, and Equations

Definition 1.1 The Pearson product-moment coefficient is a measure of the linear correlation between two variables X and Y. It has a value between the range -1 to +1, where 1 is a positive linear coefficient, -1 is a negative linear coefficient and 0 is no linear correlation. Given a pair of random variables (X, Y), the coefficient is

$$\rho_{X,Y} = \frac{cov\ (X,Y)}{\sigma_X \sigma_Y} \qquad \textbf{Eq 2.1}$$

where *cov* is the covariance; σ_X is the standard deviation of X, and σ_y is the standard deviation of Y (Pearson Correlation Coefficient, 2019).

Definition 1.2 Leave one out cross-validation is a special case of K-fold cross-validation, where a single instance from the original dataset is used as validation, and the remaining instances are used as the validation data. For linear regression, the error for leave one out cross-validation can be computed using the formula

$$\frac{1}{n}\sum \frac{(y_i - \bar{y}_i)^2}{(1 - h_{ii})^2} \qquad \textbf{Eq 2.2}$$

where h_{ii} is the ith diagonal element (Witten, et al., 2016).

Definition 1.3 Linear regression is a staple method in statistics that is used to express an outcome as a linear combination of attributes with predetermined weights:

$$x = w_0 + w_1 a_1 + w_2 a_2 + \cdots + w_k a_k \qquad \textbf{Eq 2.3}$$

where x is the real value; $a_1, a_2, \ldots a_k$ are the attribute values and $w_0, w_1, \ldots, w_k$ are the weights. The training data is used to calculate the weights. The predicted value for the first instance's real value can be written as:

$$w_0 a_0^{(1)} + w_1 a_1^{(1)} + \cdots + w_k a_k^{(1)} = \sum_{j=0}^{k} w_j a_j^{(1)} \qquad \textbf{Eq 2.4}$$

where $x^{(1)}$ is the real-valued output; $a_1^{(1)}, a_2^{(1)}, \ldots, a_k^{(1)}$ are the attribute values with the subscript indicating the first instance (Witten, et al., 2016).

Definition 1.4 Ordinary least squares (OLS) is the most common type of linear least squares formulation for approximating unknown parameters in a regression model. The method minimizes the sum of the squares of the residuals resulting in a closed-form expression for the estimated value of the unknown parameter vector β.

$$\hat{\beta} = (X^T X)^{-1} X^T \qquad \textbf{Eq 2.5}$$

where y is a vector, X is a matrix whose *ij* element is the *ith* observation of the *jth* independent variable. The estimator is unbiased and consistent if the errors have finite variance and are uncorrelated with the regressors.

$$E[x_i \varepsilon_i] = 0 \qquad \textbf{Eq 2.6}$$

where x_i is the transpose of row *i* of the matrix X (Ordinary Least Squares, 2019).

Definition 1.5 Mean square error (MSE) is used to evaluate the success of the numeric prediction. MSE is the average of the individual errors (the magnitude of the errors can be ignored).

$$MSE = \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \overline{Y}_i\right)^2 \qquad \textbf{Eq 2.7}$$

where $(Y_i - \overline{Y}_i)^2$ represents the squares of the errors; n is the number of predictions from a sample of n data points, and Y is the vector of observed values of the variable being predicted (Witten, et al., 2016).

Definition 1.6 t-distributed stochastic neighbor embedding is a nonlinear dimensionality reduction technique used for embedding high dimensional data for visualization in a low dimensional space of 2-3 dimensions (t-Distributed Stochastic Neighbor Embedding,

2019). Given a set of N high-dimensional objects x_1 , …., x_N , the algorithm computes probabilities $p_{i,j,}$ proportional to the similarity of objects x_i and x_j:

$$p_{i|j} = \frac{\exp(\frac{-||x_i - x_j||^2}{2\sigma_i^2})}{\sum_{k \neq i} \exp(\frac{-||x_i - x_k||^2}{2\sigma_i^2})} \qquad \textbf{Eq 2.8}$$

Definition 1.7 Density-based spatial clustering of applications with noise (DBSCAN) is a non-parametric algorithm used for data clustering. Given a set of data points in some space, it groups together closely packed points, marking points that lie in low-density regions as outliers (DBSCAN, 2019).

2.2.2 Pearson's Product Moment Coefficient

Pearson's product-moment coefficient is statistically significant if the p-value is less than the significance level ($\alpha = 0.05$). If the p-value is less than the significance level, the null hypothesis is to be rejected. The table below shows the R-values that categorize a strong correlation:

The NIH data is two county-wise lists of the relative frequency of language variables (emotion) and socio-economic demographic information. The two lists are converted from county-wise lists into state-wise lists. This is done by taking the column-wise mean of all the counties by state. The ever-diagnosed values obtained from the CDC website are added as the column prevalence to each of the matrices. This yields two datasets, one 50x9 matrix, and one 50x27 matrix, where the rows are the states, and the columns are emotion or socio-economic categories and prevalence.

Table 2-1: Correlation ranges and strength of the relationship.

Value of R	Strength of Relationship
-1.0 to -0.5 or 1.0 to 0.5	Strong
-0.5 to -0.3 or 0.5 to 0.3	Moderate
-0.3 to -0.1 or 0.3 to 0.1	Weak
-0.1 to 0.1	None or very weak

The two matrices are used as input to calculate the correlation values using R. R provides the *cor* method, which takes as arguments the data and the type of correlation to be performed: Pearson (default), Kendall, or Spearman. In this case, the correlation method is the Pearson correlation. The three correlation matrices are calculated for language use and ADHD prevalence ever diagnosed (referred to as the emotional prevalence in this document), socio-economic status prevalence and emotion, socio-economic status prevalence. The last correlation matrix is obtained by simply combining the state-wise matrix for emotion and socio-economic categories and running the Pearson correlation on the combined matrix.

2.2.3 Leave One Out Cross-Validation

Leave one out cross-validation (LOOCV) is a type of K-fold cross-validation, with K equaling the total number of data points in the set, N. This just means that N number of times, the function estimator is trained on all the data points except one and tested on the data point that was left out (Schneider, 1997).

The cross-validation is executed on the combined dataset: emotion + socioeconomic status. LOOCV is performed using the R package, boot. The package

provides the *glm* function that performs linear regression if the family parameter isn't passed as an argument. The function fits the model across the entire dataset. The *cv.glm* function performs the LOOCV. The result is a list of four outputs: the original call function (call), the number of folds used (K), the cross-validation estimates of prediction error (delta), and the values of the random seed used for the function call (seed).

2.2.4 Method of Least Squares Prediction, Linear Regression

The method used for the least-squares prediction is ordinary least squares (OLS). The method chooses the parameters of a linear function by minimizing the sum of the squares of the differences between the observed value and the value predicted by the linear function.

2.2.5 Metric of Success, Mean Square Error

The mean square error (MSE) is an assessment of the quality of a predictor that is more sensitive to more significant errors due to the squaring of the error. It is strictly non-negative, and lower values indicate a higher quality model.

The MSE is obtained from the result of the leave one out cross-validation. The result of the cross-validation is a list of four outputs. The first number in delta is the test error or the mean square error.

2.2.6 Multicollinearity

Multicollinearity is when a multiple regression model predictor variable can be linearly predicted from the other predictor variables. The correlation matrices for the overall models (Emotions, SES, Emotions + SES) show multicollinearity (the values above and below the diagonal are higher than 0.5). Multicollinearity in correlation matrices can be visualized using heatmaps.

Heatmaps for the three correlation matrices, the prevalence of emotion, the prevalence of socioeconomic status, and emotion + socio-economic status are created. To generate heatmaps for the correlation matrices, the packages seaborn is employed in Python.

2.2.7 t-SNE and Clustering

The matrices for emotional prevalence and socio-economic prevalence are used as input for this step. To perform t-SNE and clustering, the Scikit-learn package provided by python is used.

The two input matrices are individually transformed using the fit transform and standard scaler method provided by the package. The standard scaler method standardizes the features by removing the mean and scaling to unit variance. The standard score (z) of a training set *x* is calculated as:

$$z = \frac{(x - u)}{s} \qquad \textbf{Eq 2.9}$$

where *u* is the mean and *s* is the standard deviation of the training set. The transformed matrices are combined using the append method provided by the Numpy package, with the *axis* parameter value set to 1.

Due to the high-dimensionality of the data, t-SNE is used to visualize the data. t-SNE converts similar data points to joint probabilities, minimizing the Kullback-Leibler divergence between the probabilities of the low-dimensional embedding and the high dimensional data. Since the cost function of t-SNE is not convex, different initializations (changes in the values of the parameters) yield different results. The values for the parameters are set using experimentally determined values.

Table 2-2: Algorithm for t-SNE and DBSCAN clustering.

Input: 50x8 matrix for emotional language (X) and 50x26 matrix socio-economic status (Y).

Output: Geographical prevalence clusters and means.

1. Transform and standardize matrices X and Y.
2. Append matrices X and Y to matrix Z along axis 1.
3. Perform TSNE() on matrix Z with arguments perplexity 5, random_state 2.
4. Cluster results of step 3 with parameters eps 50 and min_samples 1.
5. **for** label in 1: unique(labels) **do**

   Create scatter plot to visualize clusters.

   end for
6. Create lists for emotion categories and prevalence *categories,* and *states.*
7. for label in 1: unique(labels) do

   print mean of emotion categories and prevalence.

   print states in clusters.

   STOP

The parameters for this step, perplexity and random state, are set to 5 and 2, respectively. DBSCAN clustering is performed on the data, and the fit predictive method is used to obtain the labeling results of running the model on the data. The parameters for clustering, eps, and min samples are set to 50 and 1. The parameter eps is the maximum distance

between the samples for a data point to be considered in the neighborhood of another data point. A scatter plot of the resulting DBSCAN labels is created to show the clusters.

To calculate the prevalence of the states, the means of the eight emotional categories and prevalence are evaluated. Two lists, categories, and states are initialized. The latter contains emotional categories and prevalence. The former is a list of the 50 states in the US. For each of the categories, the mean is calculated using the mean method provided by the Numpy package in Python, with the parameter *axis* set to 0.

2.3 Results

2.3.1 Pearson's Product Moment Coefficient

The table below shows the result of Pearson Correlation. The columns in the tables 2-3 and 2-4 from left to right are prevalence by state, anger, anxious, disengagement, engagement, negative emotion, positive emotion, negative relationship, positive relationship.

Table 2-3: Correlation values for prevalence and emotion.

Prevalence	Anger	Anx	Disegmnt	Engmnt	Neg E	Pos E	Neg R	Pos R
Ever	0.40	0.09	0.43	-0.37	-0.34	-0.25	0.48	0.34
Current	0.39	-0.09	0.38	-0.38	-0.33	-0.29	0.43	0.30
Medicated	0.39	0.14	0.42	-0.40	0.36	0.25	0.49	0.39
Medicated & Diagnosed	0.14	0.14	0.13	-0.27	0.11	0.11	0.26	0.11

Table 2-4: P-values for prevalence and emotion.

Prevalence	Anger	Anx	Disegmnt	Engmnt	Neg E	Pos E	Neg R	Pos R
Ever	0.007	0.54	0.007	0.006	0.02	0.10	0.002	0.04
Current	0.003	0.56	0.002	0.007	0.02	0.08	0.00	0.02
Medicated	0.01	0.33	0.002	0.004	0.01	0.08	0.00	0.005
Medicated & Diagnosed	0.32	0.29	0.36	0.06	0.46	0.08	0.06	0.47

The results in tables 2-3 and 2-4 show a moderate relationship between ever diagnosis, current diagnosis, medicated, and all the emotions except for anxious. The medicated and diagnosed results show a moderate relationship with engagement, positive emotion, and negative relationships. The results for the emotions anger and disengagement are positively correlated with all four groups of ADHD patients, but their correlation with medicated and diagnosed patients is weak. It is conjectured that these weak correlations imply that patients who have been medicated for ADHD are better able to control behavior that characterizes the disorder.

2.3.2 Method of Least Squares, Linear Regression

The scatter plots in figures 2-1 to 2-7 below show the results of linear regression for the features: prevalence, anger, anxious, disengagement, engagement, negative emotions, positive emotions, female population, Hispanic population, black population, foreign-born, married male, married female, high school graduate, graduate, income, smoker, diabetic, obese, fair poor health, physical unhealth days, mental unhealth days , hypertension male, hypertension female, high school/bachelor's graduate, hypertension,

married, log income, UCD, user word total, population 2010, GINI, unemployment. Four scatter plots are shown on each page. The linear equation, Pearson product-moment coefficient, P-value, t-statistic, and F-statistic for the plots is given on the right of each plot. The y-axis is the prevalence of the feature. The x-axis is the feature distribution.

The histograms in figures 2-9 to 2-15 shows the feature distribution for the emotional prevalence and the socio-economic status prevalence. Five histograms have been shown on each page. The features shown in the histogram are: prevalence, anger, anxious, disengagement, engagement, negative emotions, positive emotions, female population, Hispanic population, black population, foreign-born, married male, married female, high school graduate, graduate, income, smoker, diabetic, obese, fair poor health, physical unhealth days, mental unhealth days , hypertension male, hypertension female, high school/bachelor's graduate, hypertension, married, log income, UCD, user word total, population 2010, GINI, unemployment. Five histograms have been shown on each page. The y-axis is the frequency of the feature and the x-axis is the distribution of the feature across data points.

The feature distributions with the highest frequency are anxious, engagement, positive emotions, positive relationships, Hispanic population, black population, foreign born, married male, physical unhealth days, mental unhealth days, hypertension female, and user word total. The feature distributions with the lowest frequency are anger, negative relationships, high school graduate, bachelor’s degree, diabetic, obese, fair poor health, high school/bachelor graduate, UCD, GINI, and unemployment.

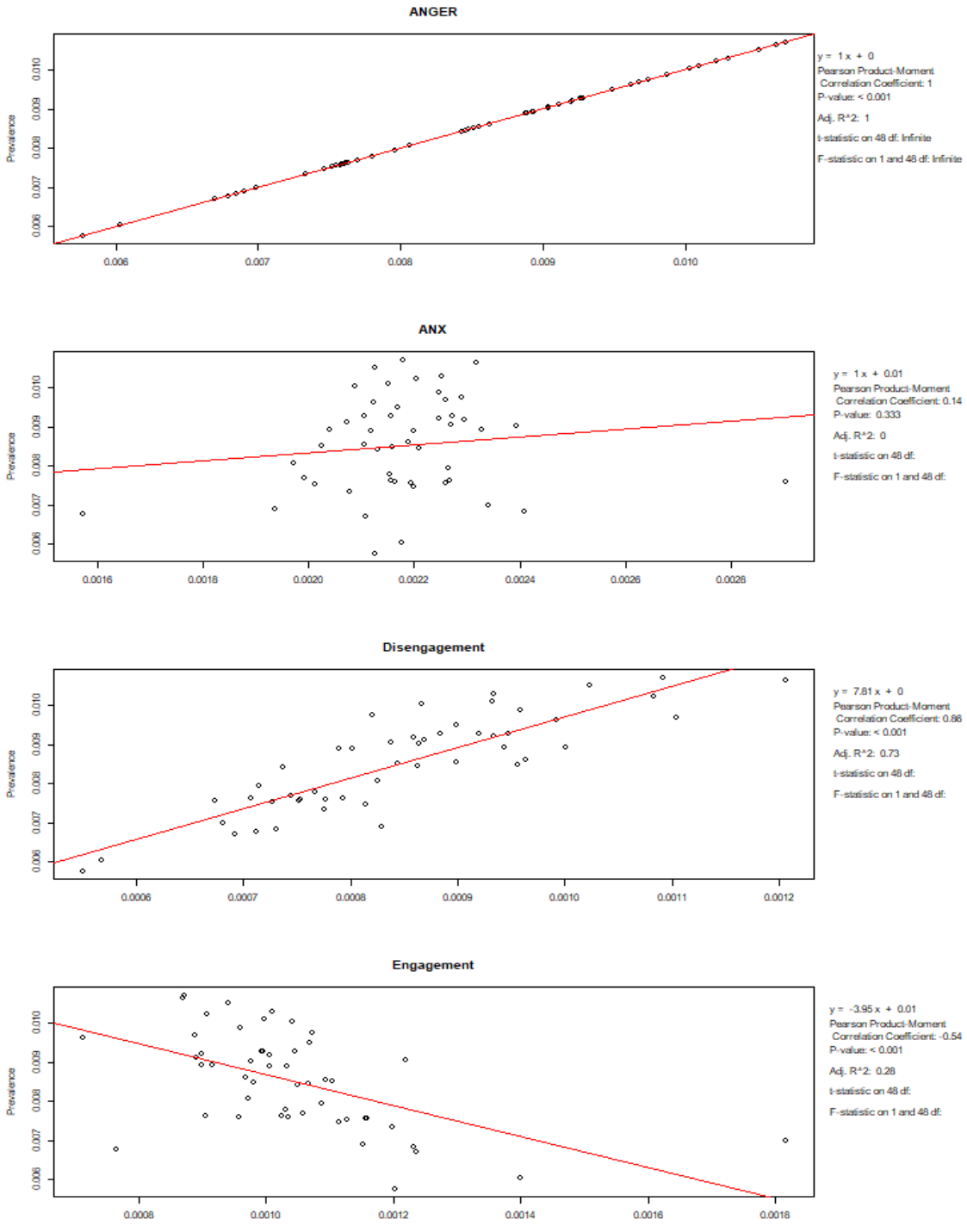

Figure 2-1: Scatter plot for feature distribution of anger, anxious, disengagement, and engagement.

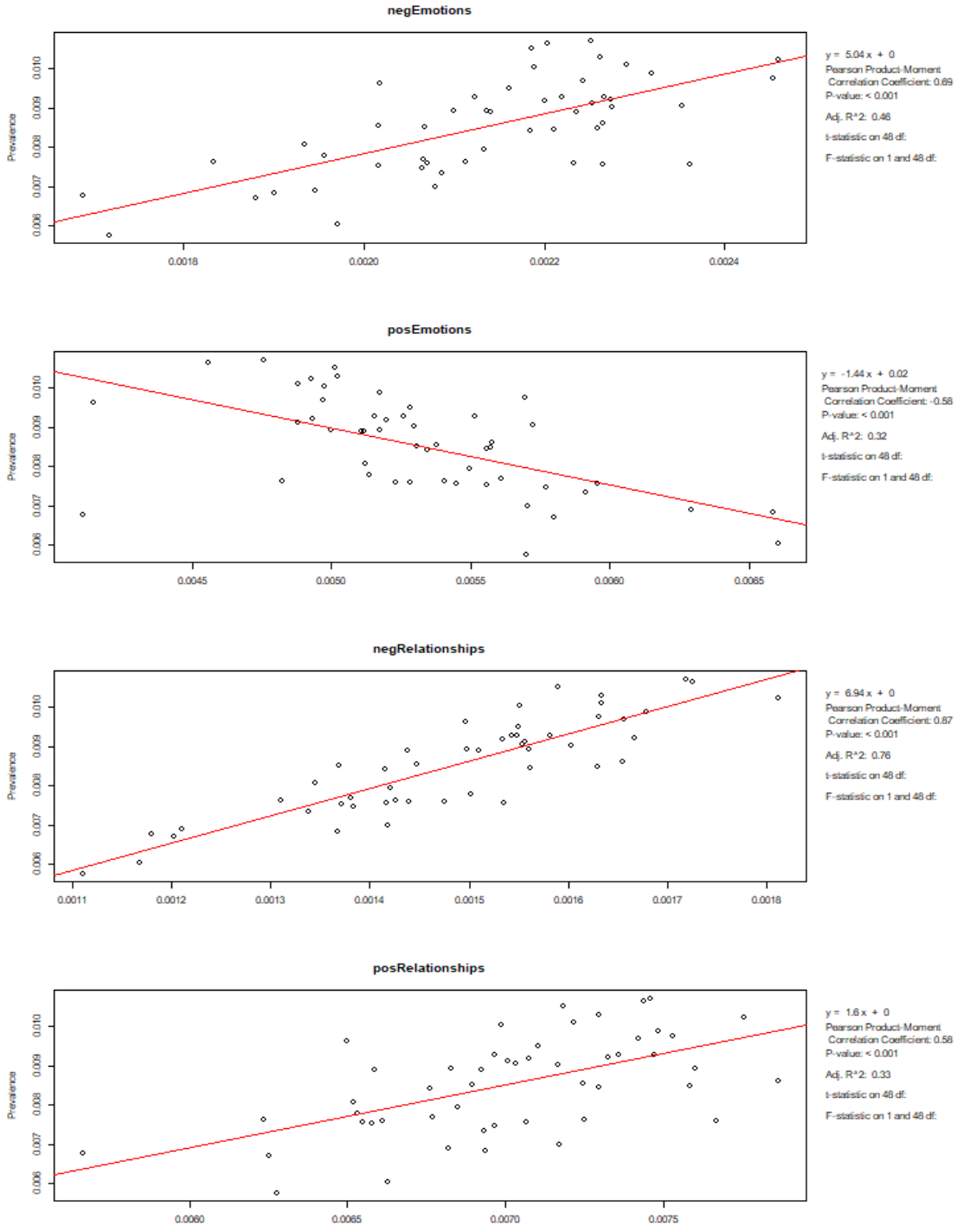

Figure 2-2: Scatter plot for feature distribution of negative emotions, positive emotions, negative relationships, positive relationships.

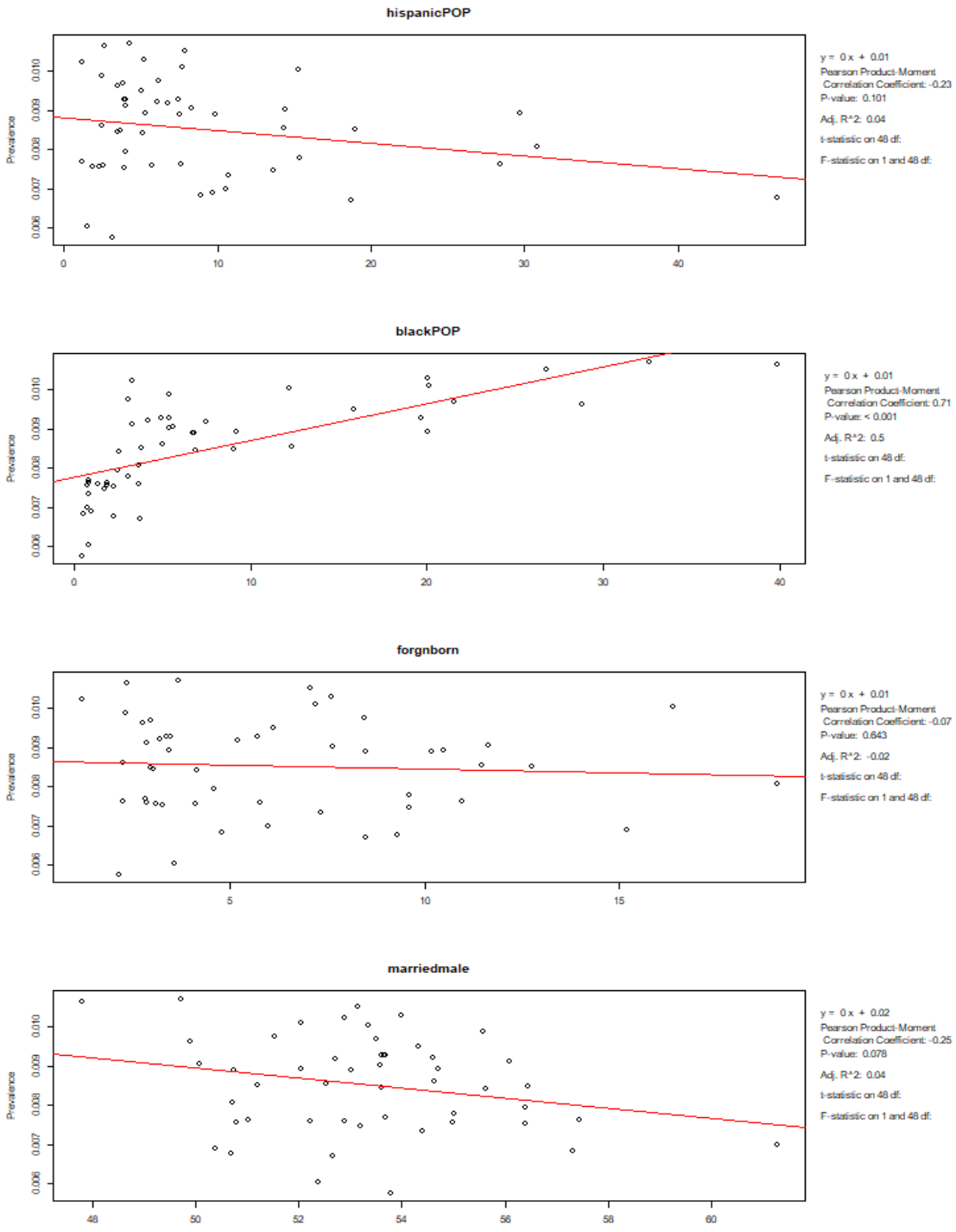

Figure 2-3: Scatter plot for feature distribution of Hispanic population, black population, foreign-born, married male.

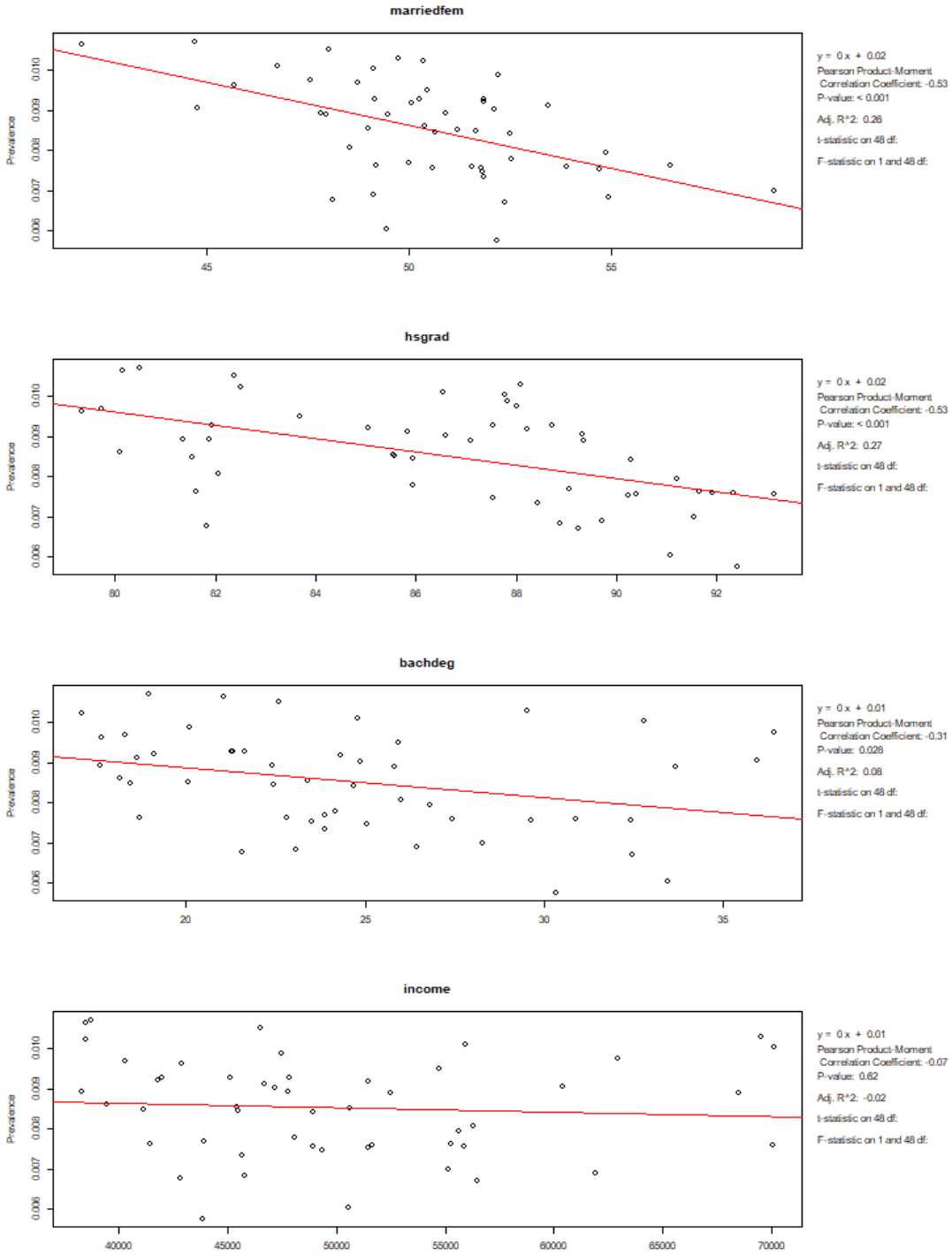

Figure 2-4: Scatter plot for feature distribution of married female, high school graduate, bachelor degree, income.

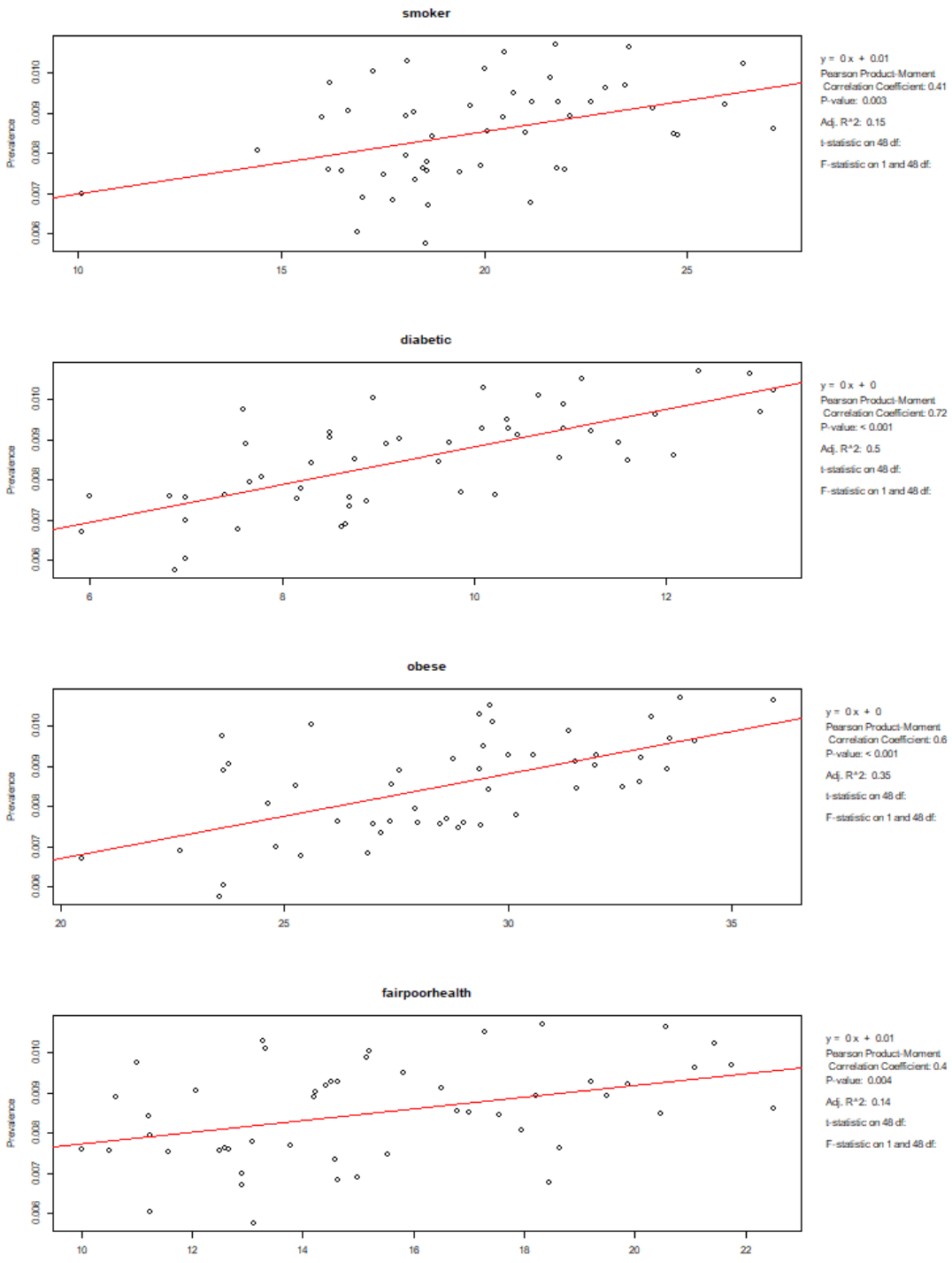

Figure 2-5: Scatter plot for feature distribution of smoker, diabetic, obese, fair poor health.

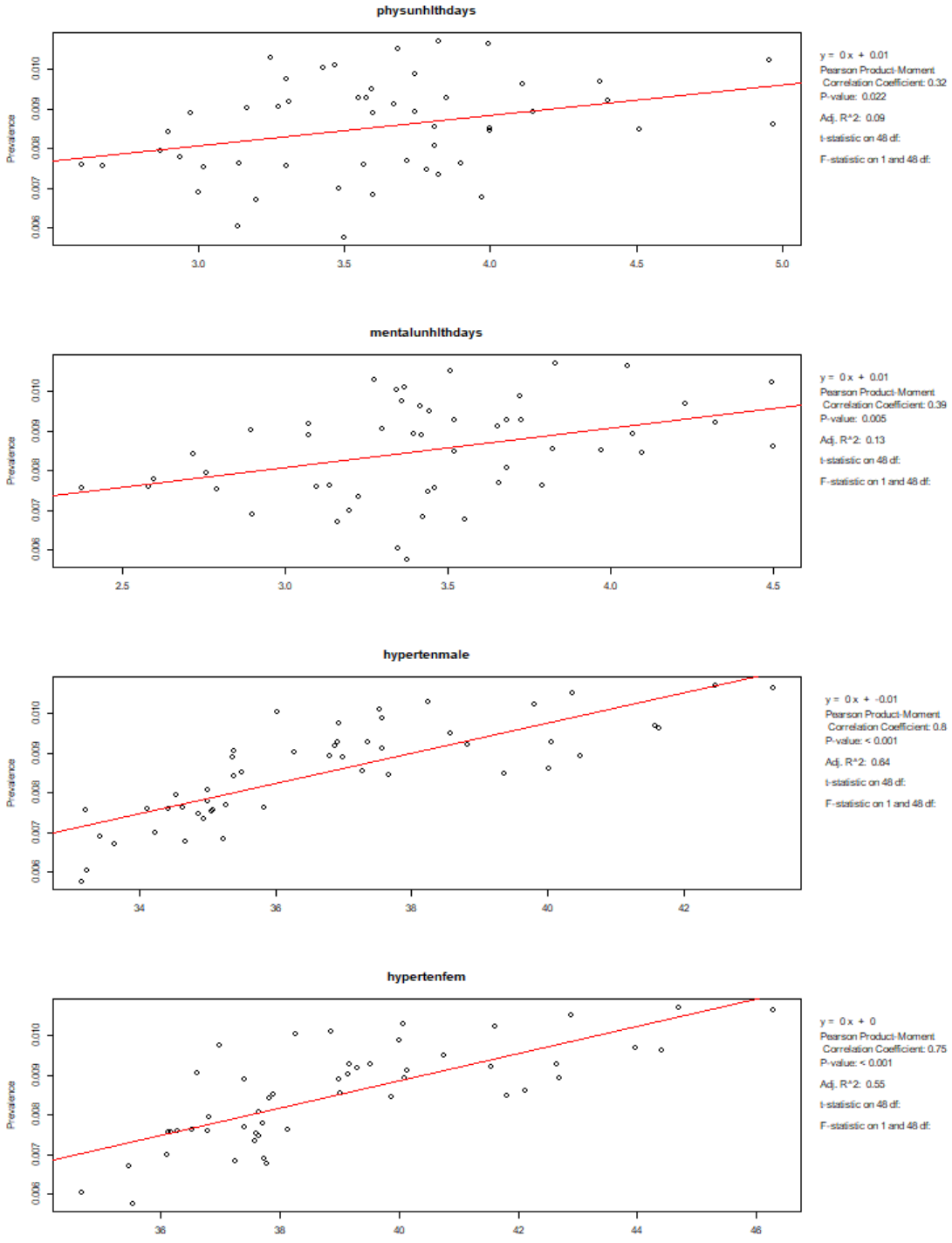

Figure 2-6: Scatter plot for feature distribution of physical unhealth days, mental unhealth days, hypertension male, hypertension female.

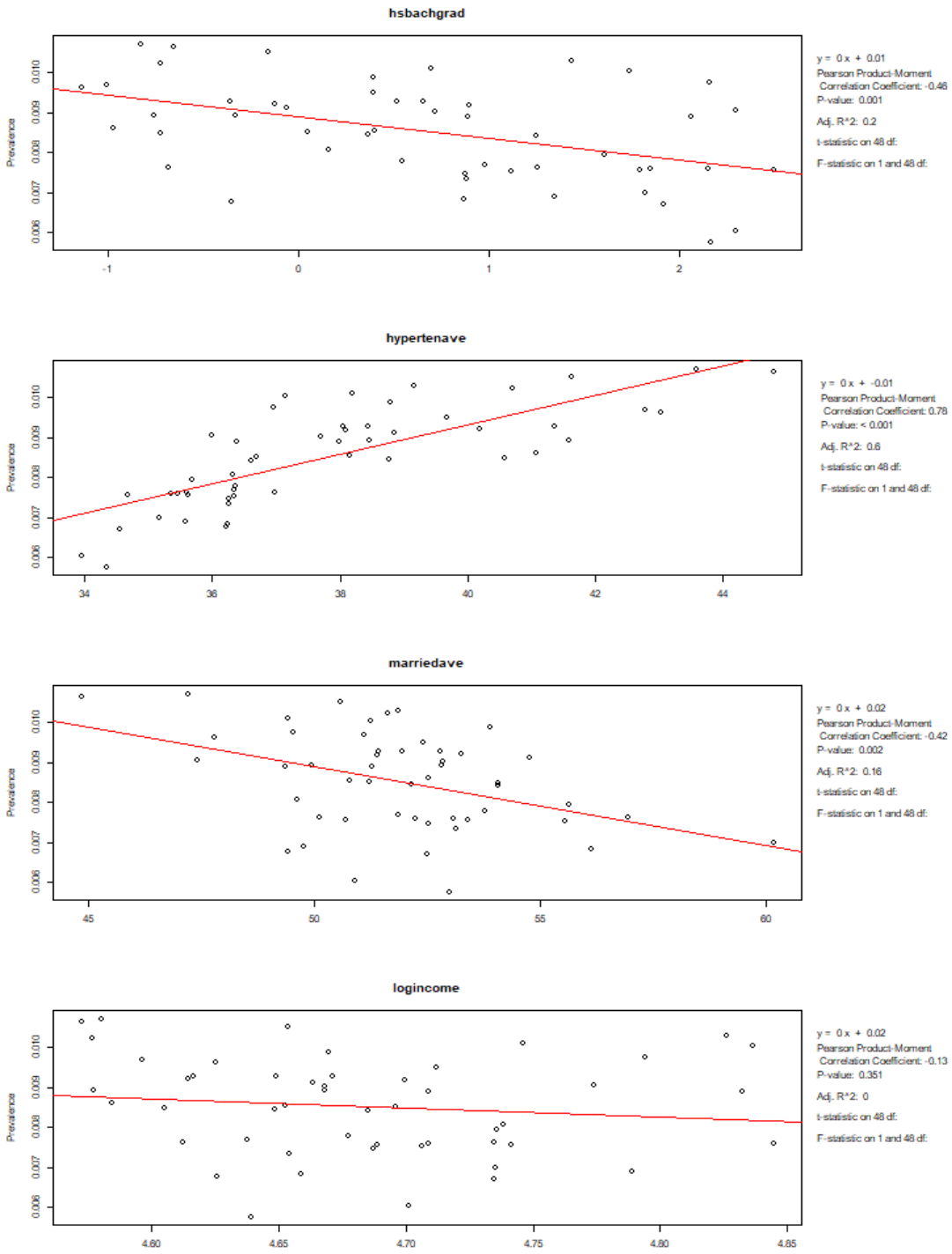

Figure 2-7: Scatter plot for feature distribution of high school/bachelor grad, hypertension, married, log income.

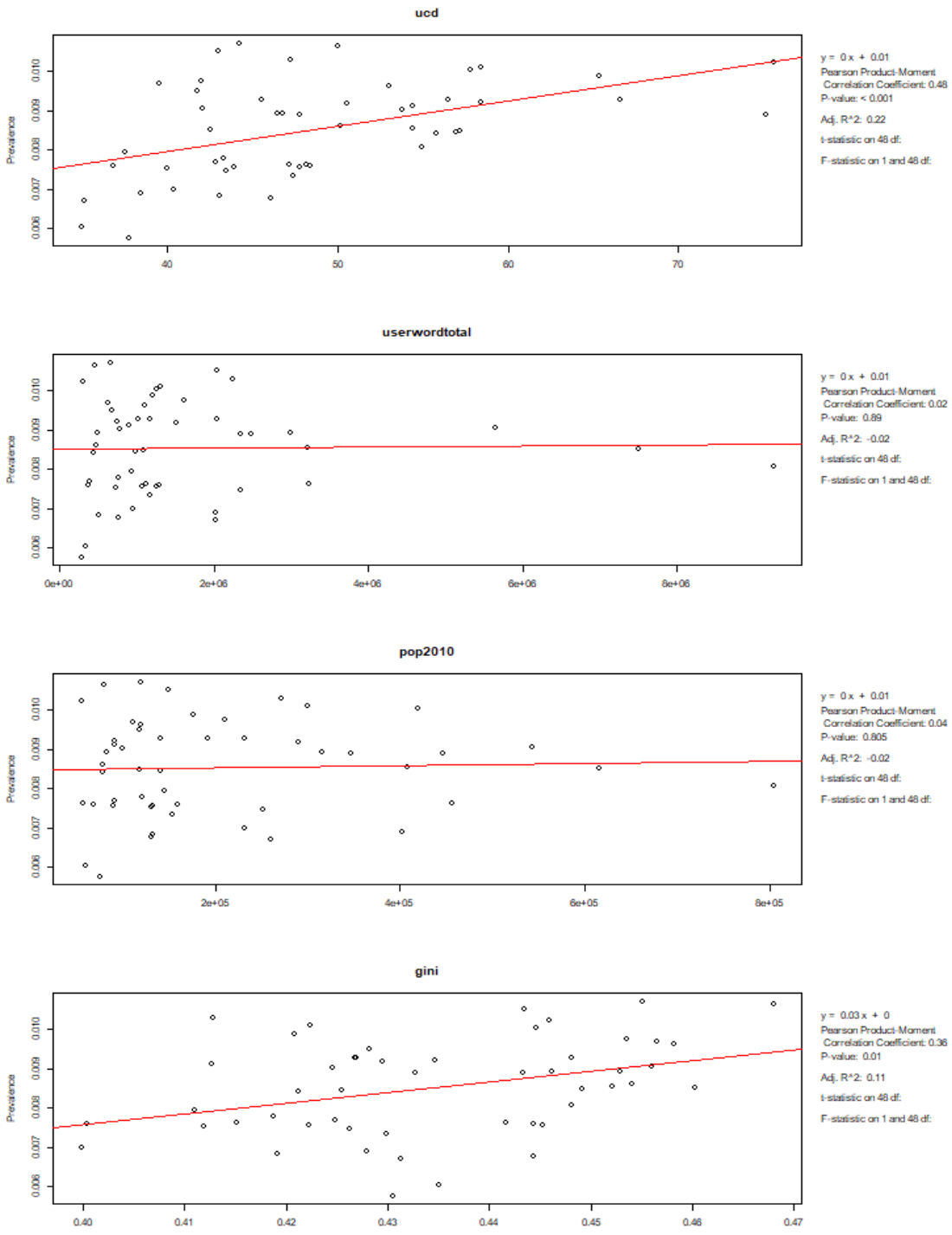

Figure 2-8: Scatter plot for feature distribution of UCD, user word total

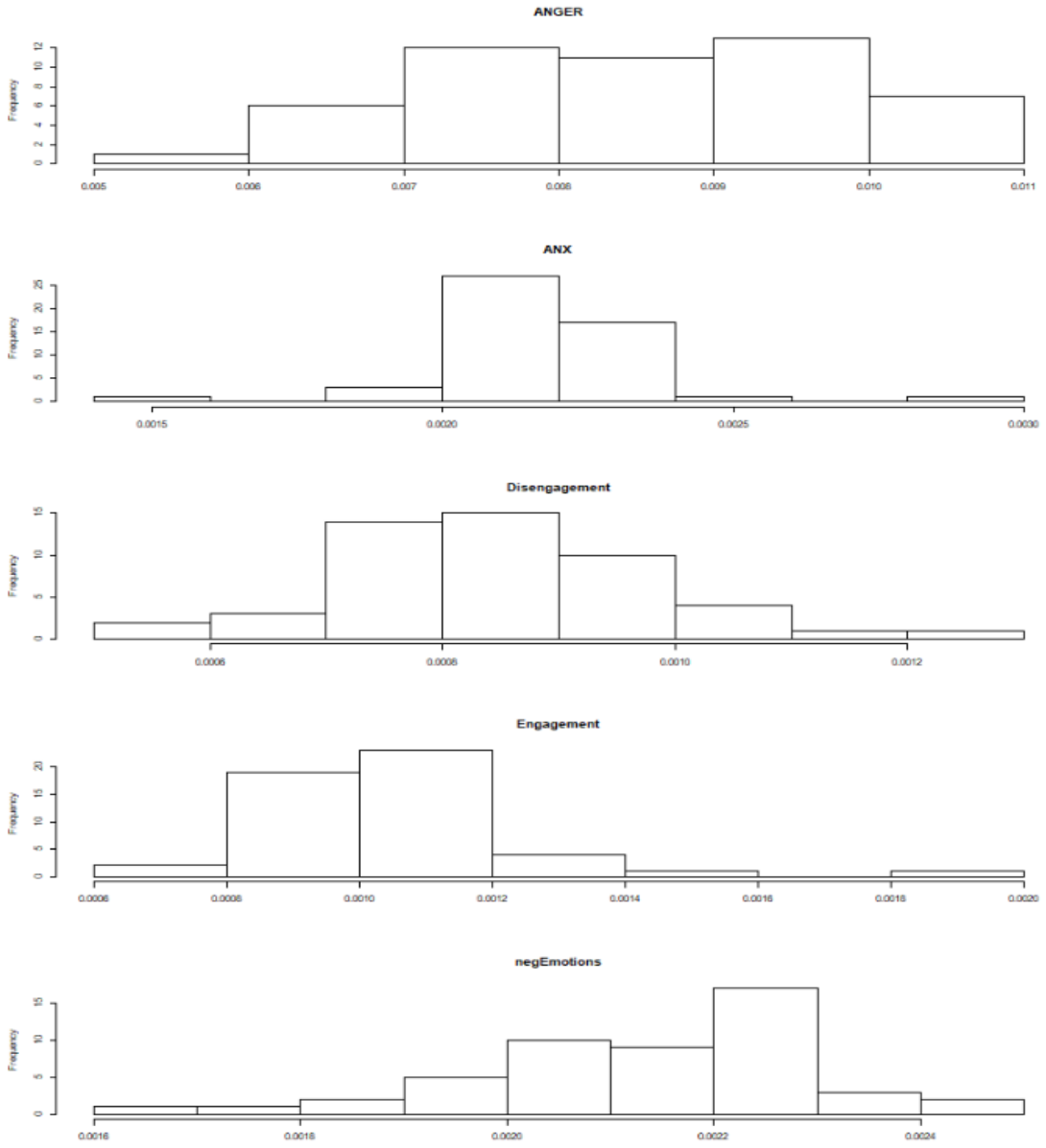

Figure 2-9: Histogram for feature distribution of anger, anxious, disengagement, engagement, and negative emotions.

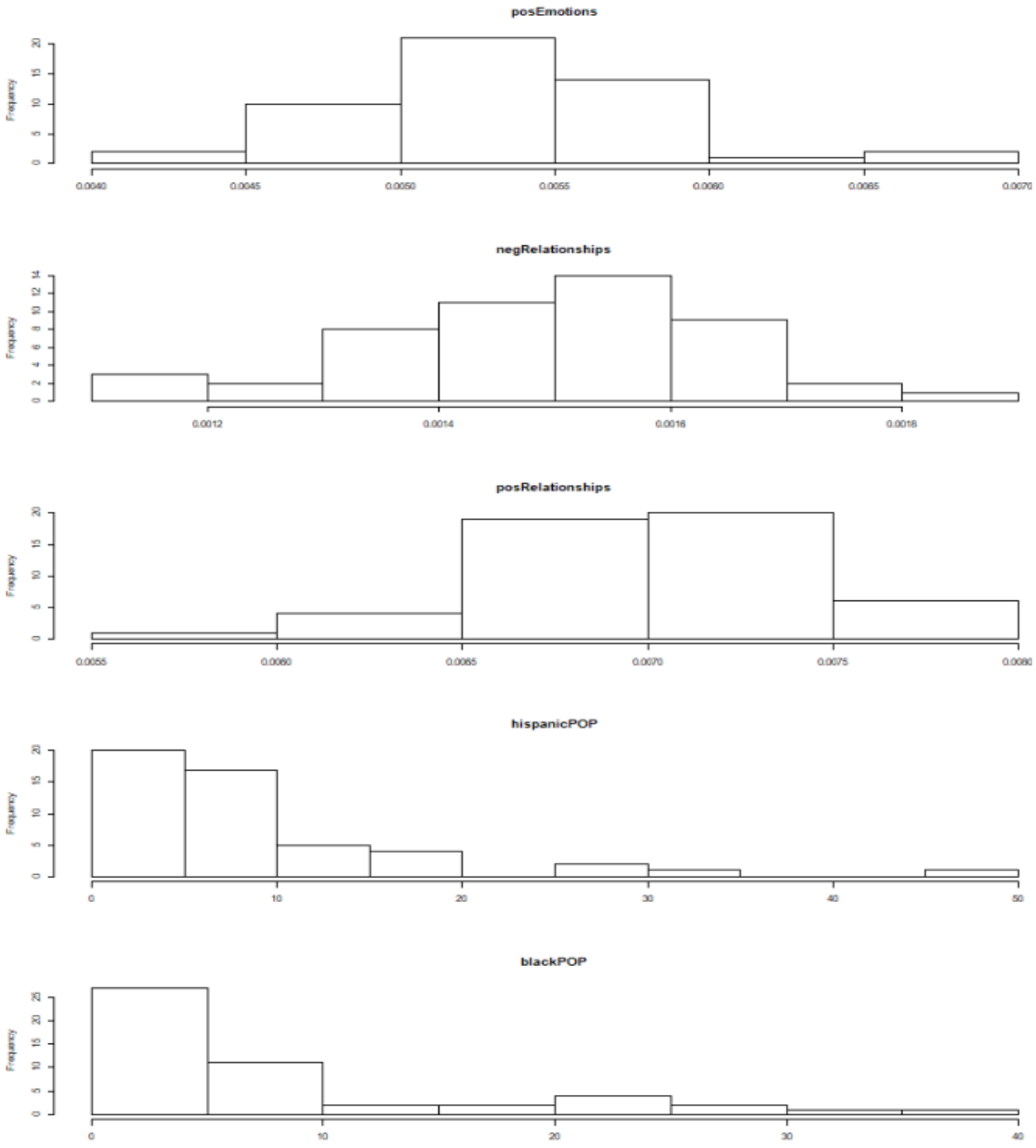

Figure 2-10: Histogram for feature distribution of positive emotion, negative relationship, positive relationship, the Hispanic population, and black population.

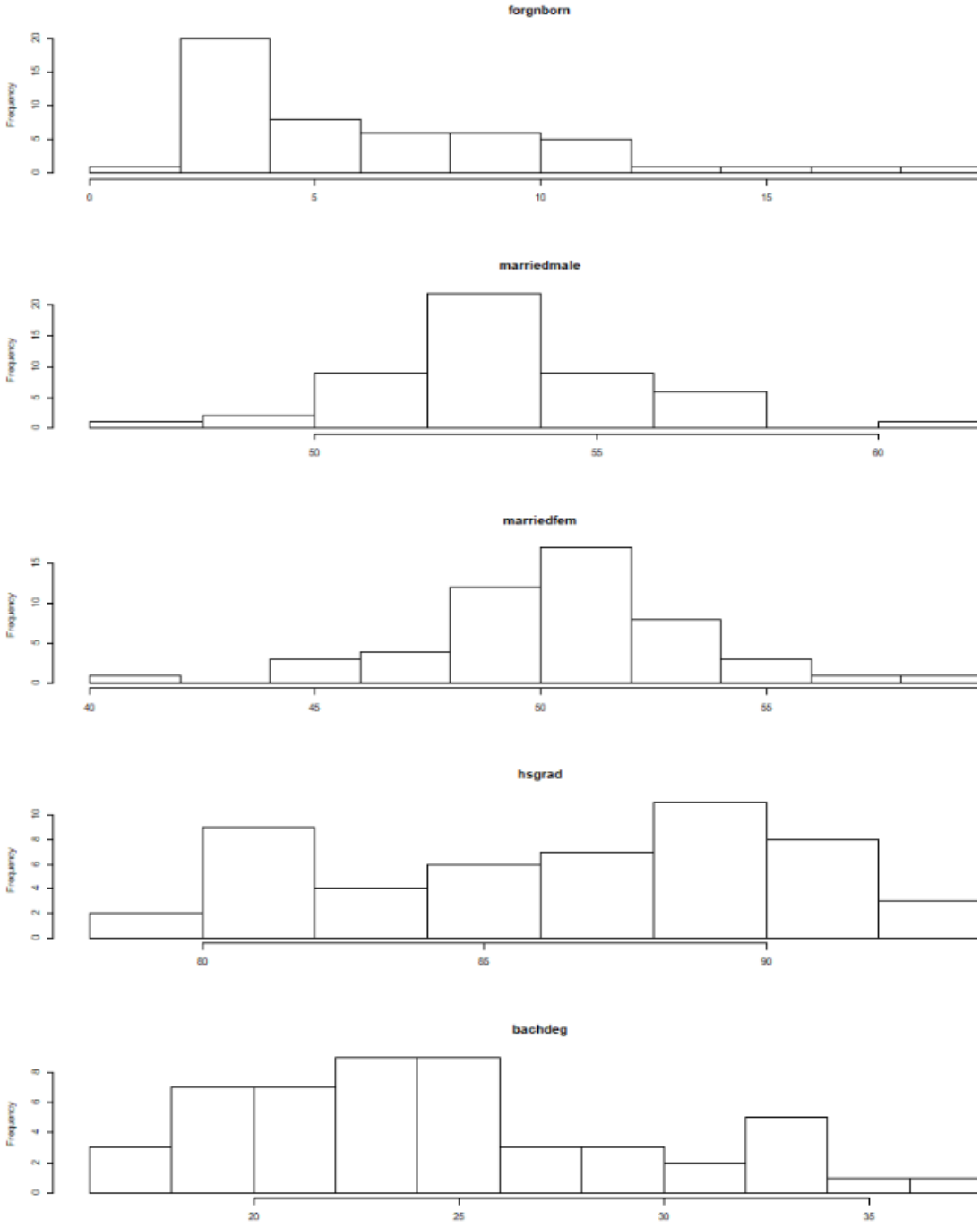

Figure 2-11: Histogram for feature distribution of foreign-born, married male, married female, high school graduate, graduate.

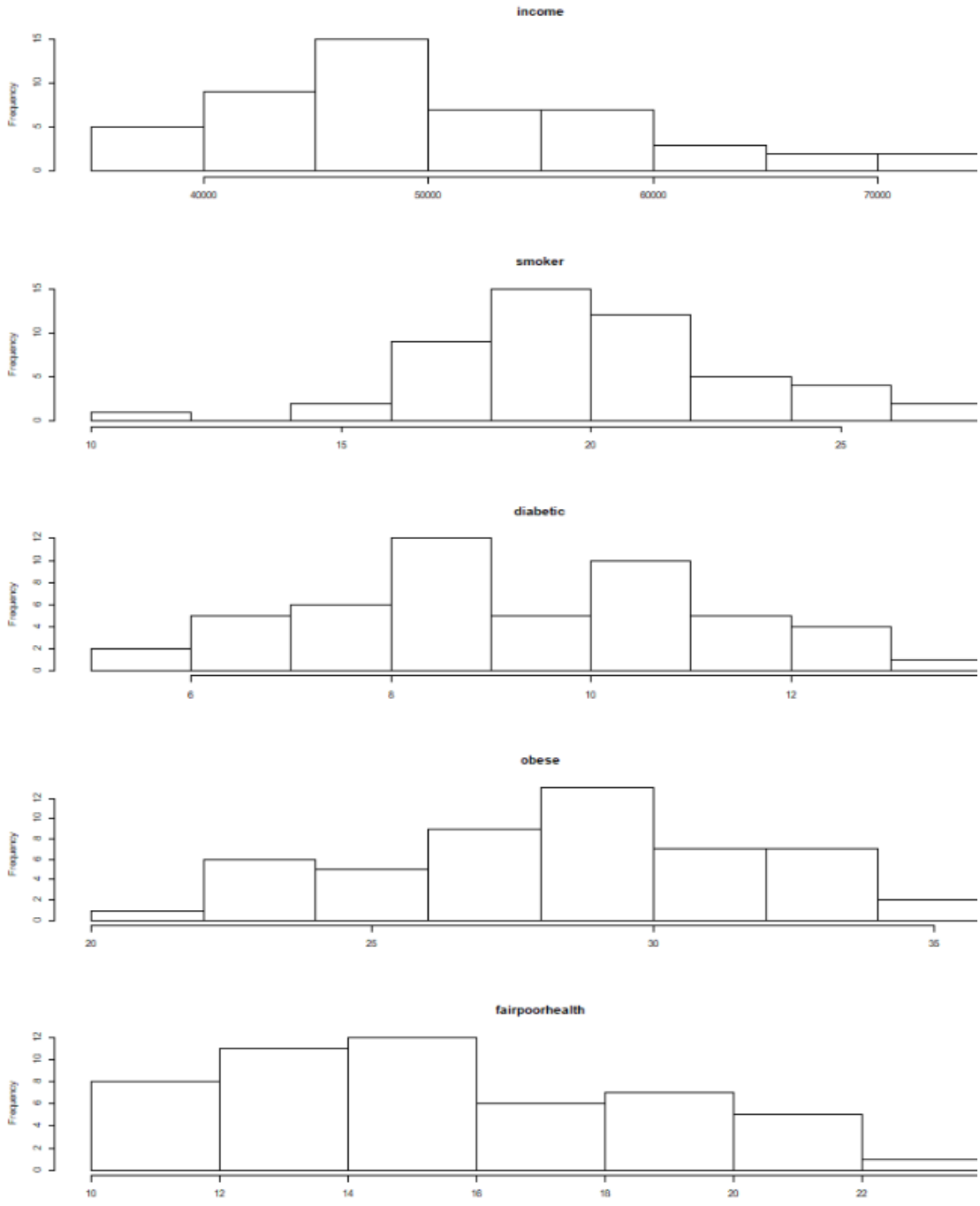

Figure 2-12: Histogram for feature distribution of income, smoker, diabetic, obese, fair, poor health.

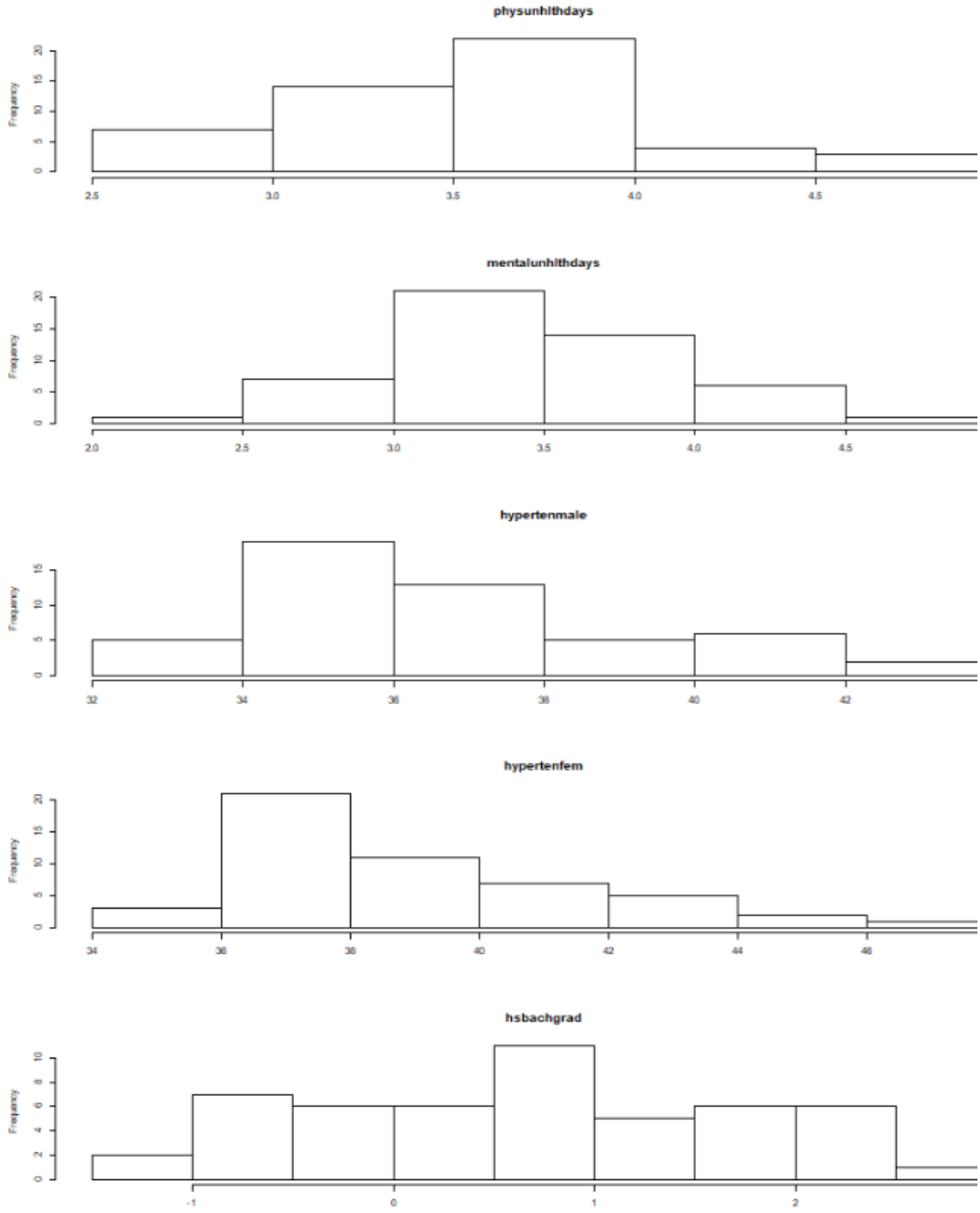

Figure 2-13: Histogram for feature distribution of physical unhealth days, mental unhealth days, hypertension male, hypertension female, high school/bachelor's graduate.

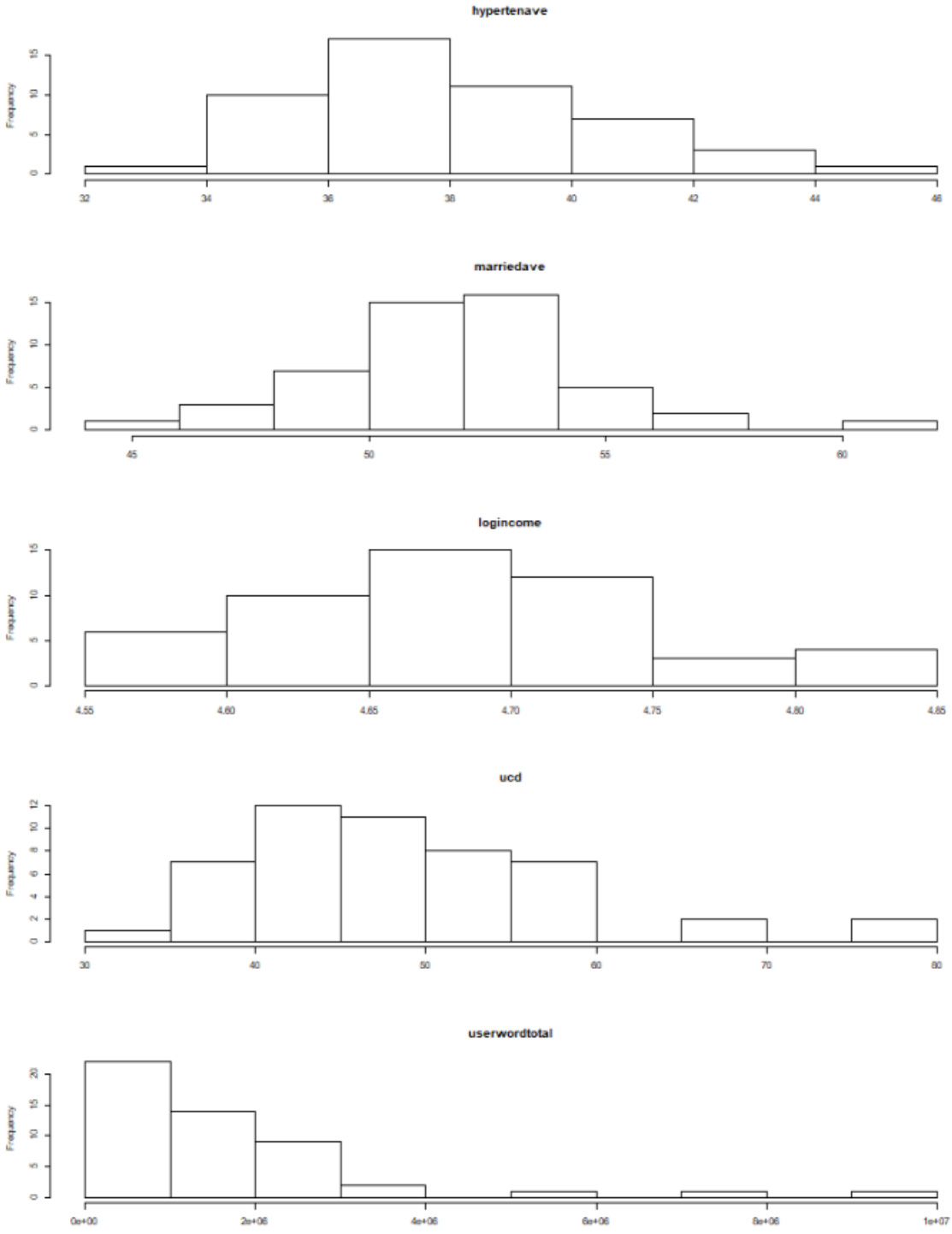

Figure 2-14: Histogram for feature distribution of hypertension, married, log income, UCD, user word total.

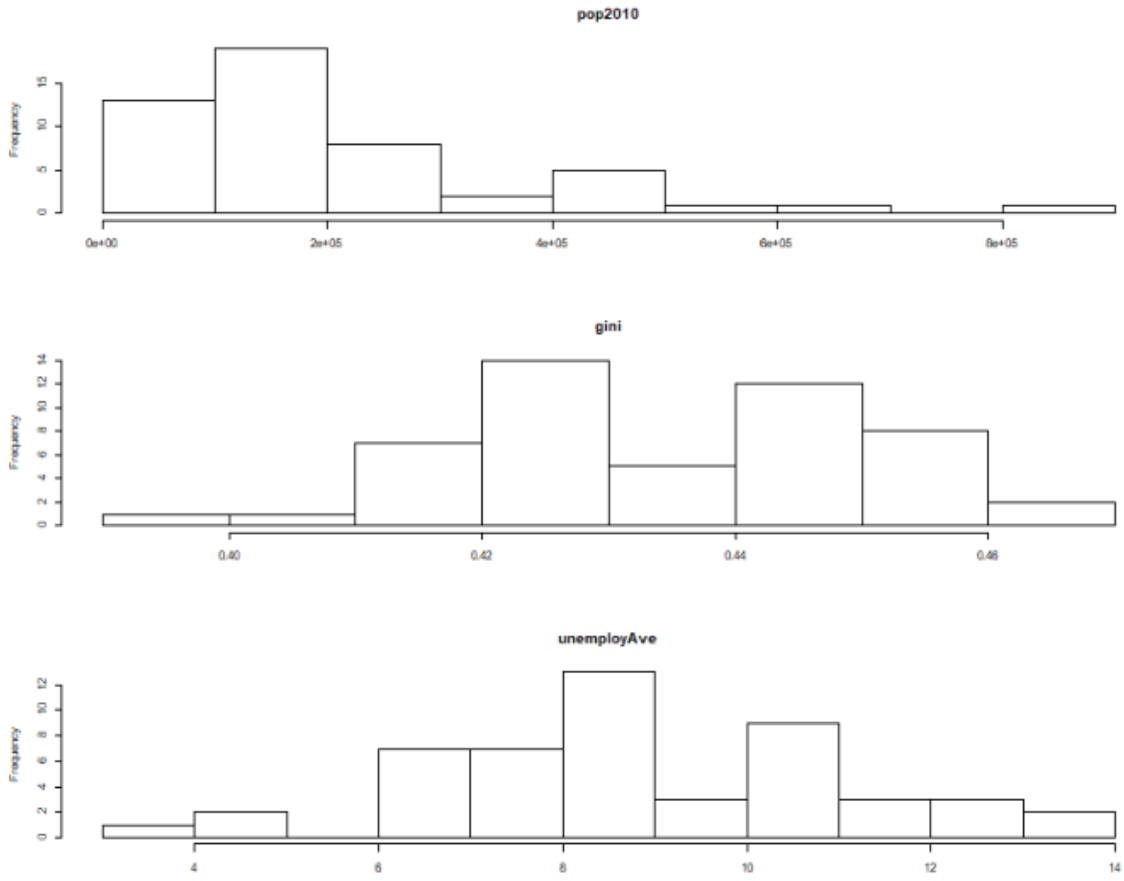

Figure 2-15: Histogram for feature distribution of population 2010, gini, unemployment.

2.3.3 Metric of Success, MSE

The table 2-5 and the bar plots (figures 2-16 and 2-17) below show the features and the MSE for the predicted prevalence and the individual features.

Table 2-5: Predicted prevalence of ADHD and emotions, SES, emotions+SES.

Feature	Mean Square Error
Emotions	9.05
SES	8.24

Table 2-5: Predicted prevalence of ADHD and emotions, SES, emotions + SES.

Feature	Mean Square Error
Emotions + SES	7.55
Anger	6.62
Anxious	8.30
Disengagement	6.46
Engagement	6.89
Negative Emotion	7.16
Positive Emotion	7.60
Negative Relationship	6.23
Positive Relationship	7.21
Female Population	6.03
Hispanic Population	6.86
Black Population	6.83
Foreign Born	6.08
Married Male	8.14
Married Female	7.33
High School Grad	6.64
Bachelor's Degree	6.73
Income	6.30
Smoker	6.22
Diabetic	4.85
Obese	5.07

Table 2-5: Predicted prevalence of ADHD and Emotions, SES, Emotions+SES.

Feature	Mean Square Error
Fair Poor Health	7.02
Physical Unhealth Days	7.17
Mental Unhealth Days	7.38
Hypertension Male	5.33
Hypertension Female	5.68
High School/Bachelor Graduate	6.39
Hypertension	5.48
Married	7.76
Log Income	6.16
UCD	7.63
User Word Total	6.86
Population 2010	6.76
Unemployment	7.92

The mean square error values for the prevalence predicted by each of the emotions show that Emotions + SES is better than the feature emotions alone. The emotions anxious, positive emotion, negative emotion, and positive relationship are all negative factors for ADHD patients. This reinforces the notion that people with ADHD have a hard time controlling their emotions. Similarly, as seen in the table for ADHD prevalence predicted by each of the socio-economic status, Emotions + SES performed better than SES alone. The features of diabetes, hypertension (male and female), obesity are risk factors for

ADHD. This implies that patients with ADHD are at risk for obesity, hypertension, and diabetes.

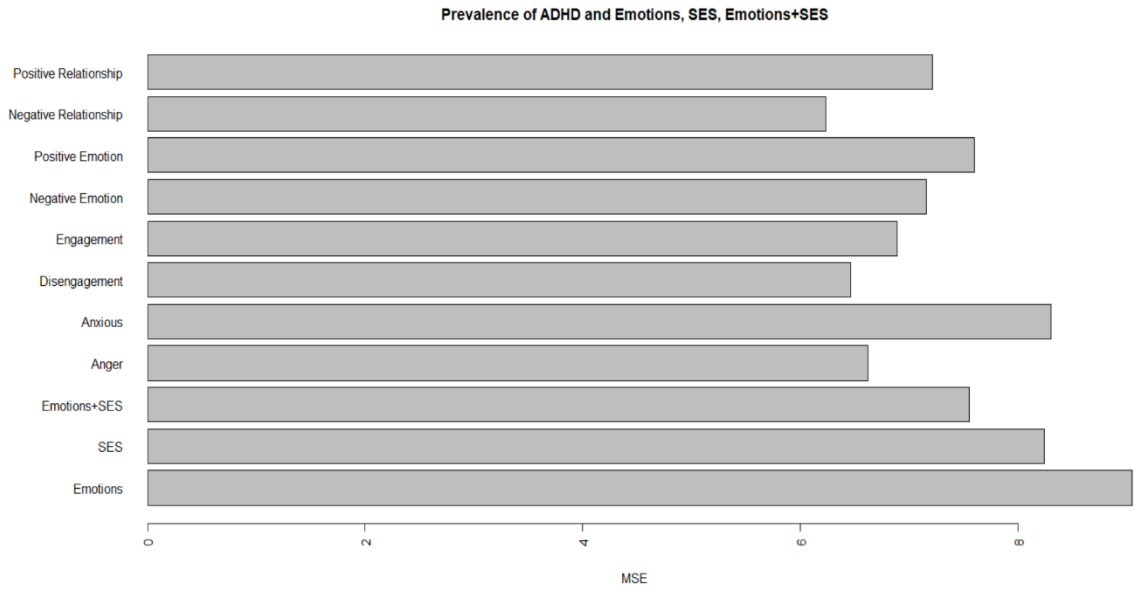

Figure 2-16: Predicted prevalence of ADHD and emotions, SES.

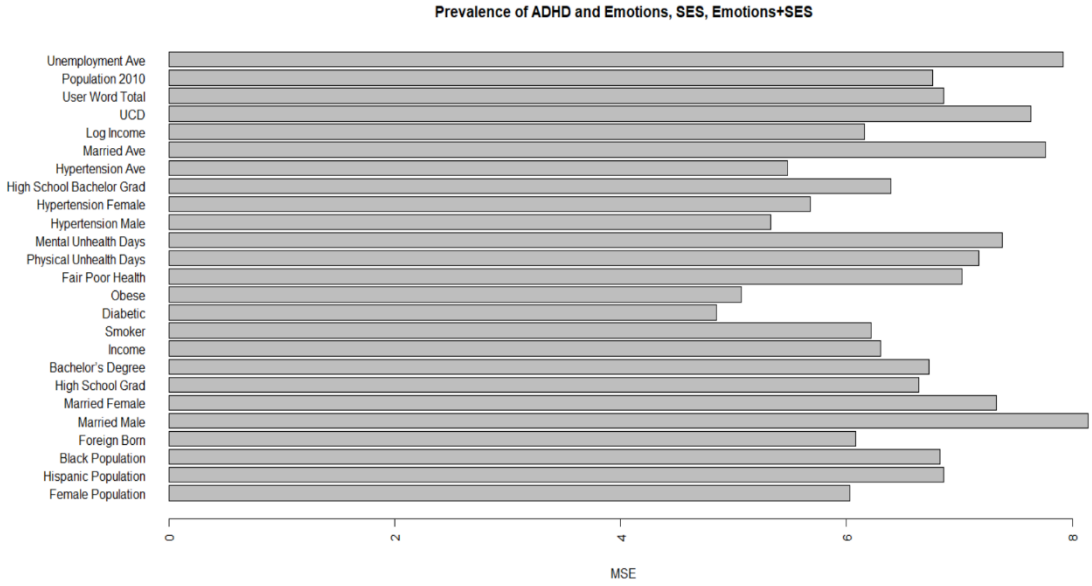

Figure 2-17: Predicted prevalence of ADHD and emotions+SES, SES.

2.3.4 Multicollinearity

Below are the heatmaps (figures 2-18 and 2-19) for the correlation matrices:

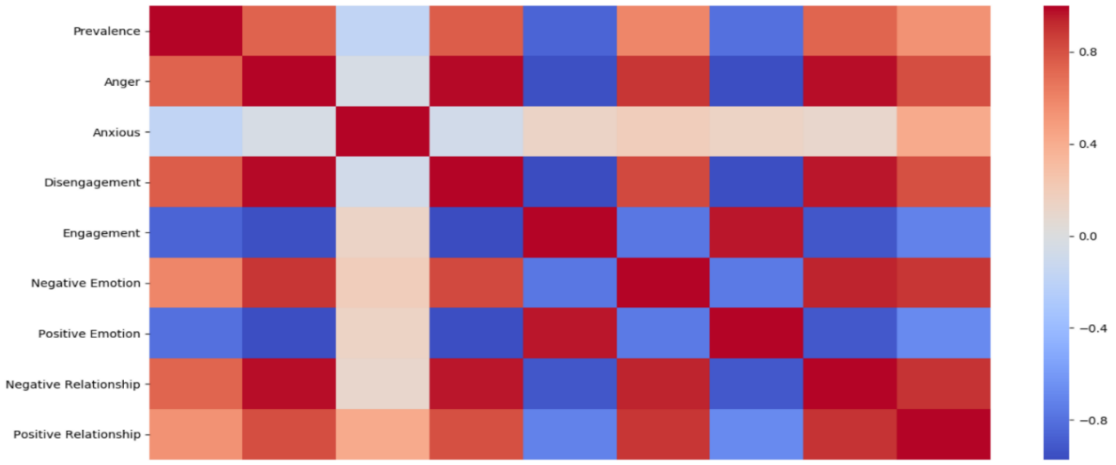

Figure 2-18: Heatmap of the correlation matrix of emotion.

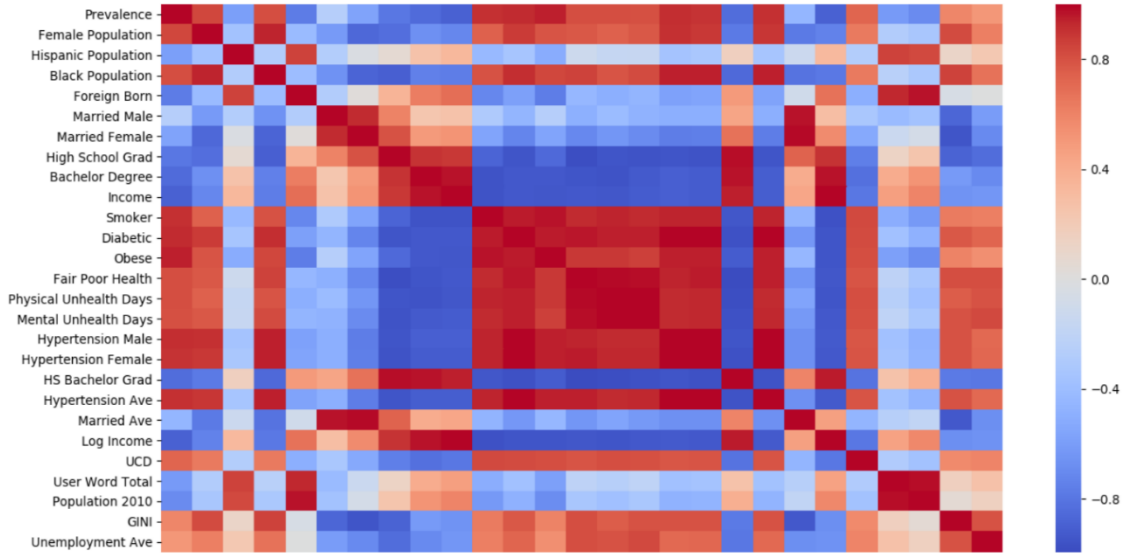

Figure 2-19: Heatmap of the correlation matrix of SES.

2.3.5 Effect Size

An effect size is a calculable measure of the value of a phenomenon. The figure 2-20 below reports the effect size of the features.

```
[1] "Highest expected prevalence  1.86X greater than lowest prevalence for ANGER"
[1] "Highest expected prevalence  1.17X greater than lowest prevalence for ANX"
[1] "Highest expected prevalence  1.83X greater than lowest prevalence for Disengagement"
[1] "Highest expected prevalence  1.80X greater than lowest prevalence for Engagement"
[1] "Highest expected prevalence  1.62X greater than lowest prevalence for negEmotions"
[1] "Highest expected prevalence  1.54X greater than lowest prevalence for posEmotions"
[1] "Highest expected prevalence  1.82X greater than lowest prevalence for negRelationships"
[1] "Highest expected prevalence  1.55X greater than lowest prevalence for posRelationships"
[1] "Highest expected prevalence  1.20X greater than lowest prevalence for hispanicPOP"
[1] "Highest expected prevalence  1.47X greater than lowest prevalence for blackPOP"
[1] "Highest expected prevalence  1.04X greater than lowest prevalence for forgnborn"
[1] "Highest expected prevalence  1.23X greater than lowest prevalence for marriedmale"
[1] "Highest expected prevalence  1.54X greater than lowest prevalence for marriedfem"
[1] "Highest expected prevalence  1.31X greater than lowest prevalence for hsgrad"
[1] "Highest expected prevalence  1.19X greater than lowest prevalence for bachdeg"
[1] "Highest expected prevalence  1.04X greater than lowest prevalence for income"
[1] "Highest expected prevalence  1.38X greater than lowest prevalence for smoker"
[1] "Highest expected prevalence  1.49X greater than lowest prevalence for diabetic"
[1] "Highest expected prevalence  1.48X greater than lowest prevalence for obese"
[1] "Highest expected prevalence  1.24X greater than lowest prevalence for fairpoorhealth"
[1] "Highest expected prevalence  1.23X greater than lowest prevalence for physunhlthdays"
[1] "Highest expected prevalence  1.28X greater than lowest prevalence for mentalunhlthdays"
[1] "Highest expected prevalence  1.54X greater than lowest prevalence for hypertenmale"
[1] "Highest expected prevalence  1.57X greater than lowest prevalence for hypertenfem"
[1] "Highest expected prevalence  1.26X greater than lowest prevalence for hsbachgrad"
[1] "Highest expected prevalence  1.56X greater than lowest prevalence for hypertenave"
[1] "Highest expected prevalence  1.44X greater than lowest prevalence for marriedave"
[1] "Highest expected prevalence  1.08X greater than lowest prevalence for logincome"
[1] "Highest expected prevalence  1.34X greater than lowest prevalence for ucd"
[1] "Highest expected prevalence  1.02X greater than lowest prevalence for userwordtotal"
[1] "Highest expected prevalence  1.02X greater than lowest prevalence for pop2010"
[1] "Highest expected prevalence  1.24X greater than lowest prevalence for gini"
[1] "Highest expected prevalence  1.25X greater than lowest prevalence for unemployAve"
```

Figure 2-20: Effect size of features

2.3.6 t-SNE and DBSCAN

A scatter plot (figure 2-21) is used to show the clusters obtained from DBSCAN. The categories in the cluster are prevalence, anger, anxious, disengagement, engagement, negative emotions, positive emotions, negative relationships, and positive relationships. Each of the data points in the clusters represents one of the 50 states.

The states in cluster 0 (represented by the color purple) are Alabama, Arkansas, Georgia, Kentucky, Louisiana, Mississippi, North Carolina, South Carolina, and Tennessee. The means for cluster 0 is 14.4, 9.6×10^{-3} , 2.20×10^{-3} 1.032×10^{-3}, 9.042×10^{-4}, 2.19×10^{-3}, 5.00×10^{-3}, 1.62×10^{-3}, 7.39×10^{-3}. These states have the highest prevalence, with a mean of 14.4% that were computationally organized by t-SNE and DBSCAN.

The states in cluster 1 (represented by the color blue) are Alaska, Colorado, Hawaii, Idaho, Iowa, Minnesota, Montana, Nebraska, New Hampshire, North Dakota, Oregon, South Dakota, Utah, Vermont, Washington, Wisconsin, and Wyoming. The means for cluster 1 is 10.34, 7.27×10^{-3}, 2.22×10^{-3}, 7.20×10^{-4}, 1.18×10^{-3}, 2.058×10^{-3}, 5.72×10^{-3}, 1.36×10^{-3}, 6.82×10^{-3}.

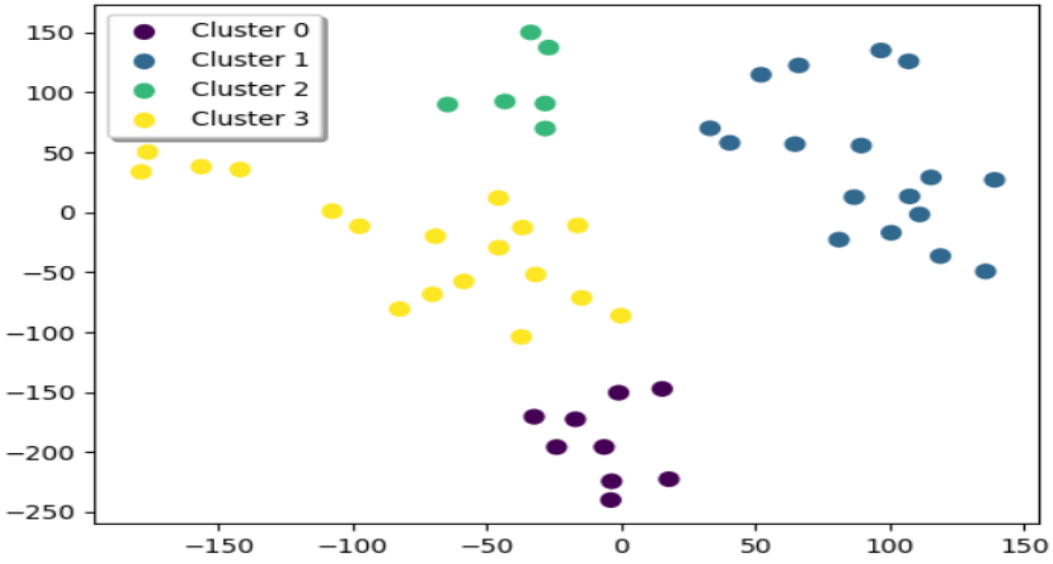

Figure 2-21: Scatter plot for the clusters obtained from DBSCAN

The states in cluster 2 (represented by the color green) are Arizona, California, Florida, Nevada, New Mexico, and Texas. The means for cluster 2 is 9.78 ,8.07×10^{-3}, 1.98×10^{-3}, 8.36×10^{-4}, 9.57×10^{-4}, 1.94×10^{-3}, 4.96×10^{-3}, 1.36×10^{-3}, 6.57×10^{-3}. These states have the lowest prevalence, with a mean of 9.8% that were computationally organized by t-SNE and DBSCAN.

The states in cluster 3 (represented by the color yellow) are Connecticut, Delaware, Illinois, Indiana, Kansas, Maine, Maryland, Massachusetts, Michigan, Missouri, New Jersey, New York, Ohio, Oklahoma, Pennsylvania, Rhode Island, Virginia, and West

Virginia. The means in cluster 3 is 11.7, 9.32×10^{-3}, 2.19×10^{-3}, 8.81×10^{-4}, 1.01×10^{-3}, 2.26×10^{-3}, 5.22×10^{-3}, 1.58×10^{-3}, 7.16×10^{-3}.

2.4 Conclusion

The results successfully establish a correlation between emotions, language use, and the prevalence of ADHD geographically in the United States. The combination of emotions and socio-economic statuses successively outperforms individual result sets. The result set could be further fortified by analyzing the prevalence of ADHD geographically by a new feature age. The nature of the behavioral disorder is such that it statistically manifests in adolescence and peaks/subsides as user ages. The age, along with other socio-economic factors, would further assist in identifying measures of the disorder.

CHAPTER 3

MEASURES OF BEHAVIORAL DISORDERS

Behavioral disorders are deficits in adults and children characterized by learning disabilities and an inability to build or maintain satisfactory interpersonal relationships (Emotional and Behavioral Disorder, 2019). Diagnosing such disorders requires the study of behavior, making it difficult for medical professionals to diagnose them. Numerous studies so far have categorized behavior into language use, social expressions, and interaction.

Attention Deficit Hyperactivity Disorder (ADHD) is a behavioral disorder characterized by significant problems with attention, impulsiveness, and hyperactivity (Attention-deficit/hyperactivity disorder, 2019). The commonplace nature of the disorder and the longstanding societal stigma associated with it leaves many more cases undiagnosed. In addition, the lack of data to efficiently diagnose the disorder has proved burdensome in providing effective treatment.

This chapter identifies two behavioral measures and an analysis of ADHD, namely, variations in phrase structure rules, topic detection, and sentiment analysis that can be further utilized in the development of a social media-based clinical decision support system to effectively aid in the diagnosis of users with a predisposition for ADHD. The chapter is organized as follows: related works, methodology, results, and conclusion.

3.1 Related Works

Social media data offers many advantages, many of which lie in the diversity in the language styles used. The diversity of language on Twitter exceeds the formal genres for the English language, such as the Penn Treebank and the Brown Corpus, mainly because there are fewer rules to follow, the more significant number of authors, and varied communicative settings (Balusu, et al., 2018). These authors worked on quantifying the impact of one form of socio-linguistic variation on the accuracy of part of speech tags. Meftah, et al. (2018) worked on a POS tagger for social media datasets, using an end-to-end neural model based on Transfer Learning. Kilyeni (2014) explored the use of '*buzzwords*' that were coined on social media platforms and are now used in daily life (on and off social media). Similarly, Qadir, et al. (2015) presented a semantic lexicon induction approach to learn new vocabulary from social media.

Surian, et al. (2016) used topic modeling methods to measure how information disseminates in online communities to effectively find the geographical variations in decisions that result in poor health outcomes. Lu, et al. (2013) integrated medical-domain specific features to analyze messages posted in online health communities.

The informal manner in which tweets are posted makes it ideal for sentiment analysis. Wang, et al. (2011) performed a hashtag level sentiment classification to analyze the overall sentiment polarity for a given period. Carchiolo, et al. (2015) exploited SNOMED-CT terminology to analyze how a disease is perceived by the public. Ji, et al. (2015) tracked an outbreak by analyzing how concerned the general population in an area

was. Researchers also determined whether the use of sentiment words of a user with depression differed from the general population.

3.2 Methodology

3.2.1 Definitions, Algorithms, and Methodology

Definition 3.2.1 Parts of speech are categories to which words are assigned in accordance with their syntactic function. The main parts of speech are noun, pronoun, adjective, determiner, verb, adverb, preposition, conjunction, and interjection.

Definition 3.2.2 Cosine similarity is a measure of the similarity between two non-zero vectors of an inner product space that measures the cosine of the angle between them. The cosine of 0 degrees is 1, and it is less than 1 for any angle in the interval (0, τ] radians. (Han, et al., 2000) The cosine of the two non-zero vectors, A and B, can be derived using the formula:

$$similarity = \cos(\theta) = \frac{A.B}{||A||\,||B||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \qquad \textbf{Eq 3.1}$$

Definition 3.2.3 Recurrent neural networks are a class of neural networks that allow previous outputs to be used as inputs while having hidden states (Han, et al., 2000). For each time step *t*, the activation $a^{<t>}$ and the output $y^{<t>}$:

$$a^{<t>} = g_1\,(W_{aa}a^{<t-1>} + W_{ax}x^{<t>} + b_a) \qquad \textbf{Eq 3.2}$$

and,

$$y^{<t>} = g_2\,(W_{ya}a^{<t>} + b_y) \qquad \textbf{Eq 3.3}$$

Definition 3.2.4 Term frequency-inverse document frequency (TF-IDF) is a numeral statistic that reflects the importance of a word for a document in a corpus (TF-IDF, 2019). It is a commonly used statistic in information retrieval, text mining, and user modeling.

The value of TF-IDF is proportional to the word count in a document and is offset by how many documents in a corpus contain the word. TF-IDF is calculated using the formula:

$$tfidf(t,d,D) = tf(t,d).idf(t,D) \quad \textbf{Eq 3.4}$$

where term is represented by *t*, document by *d* and document corpus by *D*. The term frequency *(tf)* and inverse document frequency *(idf)* are calculated as:

$$tf(t,d) = 0.5 + 0.5.\frac{f_{t,d}}{\max\{f_{t',d} : t' \in d\}} \quad \textbf{Eq 3.5}$$

where $f_{t,d}$ denotes the raw count.

$$idf\ (t,D) = \log\frac{N}{|\{d \in D: t \in d\}|} \quad \textbf{Eq 3.6}$$

where N denotes the total number of documents in the corpus.

Definition 3.2.5 Non- negative matrix factorization (NMF) is a collection of algorithms in the multivariate analysis where a matrix is factorized into two matrices W and H with the condition that all three matrices must have no negative elements.

Definition 3.2.6 Kullback-Leibler divergence is a measure of how two probability distributions differ from one another (Kullback-Leibler Divergence, 2019). For two probability distributions P and Q on the same space, the divergence is calculated as:

$$D_{KL}(P||Q) = -\sum_{x\in X}\log\frac{P(x)}{Q(x)}\ P(x) \quad \textbf{Eq 3.7}$$

which is equivalent to,

$$D_{KL}\ (P||Q) = \sum_{x\in X} P(x)\log(\frac{P(x)}{Q(x)}) \quad \textbf{Eq 3.8}$$

3.2.2 Data Collection

The data to identify behavioral measures of ADHD is collected from Twitter using their developer API. Data is collected for two groups of users: diagnosed and control. The diagnosed group is composed of users with tweets of self-reported diagnosis of the disorder. Alternatively, the control group is composed of users who have no tweets of self-reported diagnosis of the disorder.

The process used to collect the data is similar to the one used in Coppersmith, et al. (2014). The process has been previously validated and shows predictive power for real-world phenomena. For the diagnosed group, self-reported diagnosis tweets are posts containing statements such as ‘I have been diagnosed with ADHD' or ‘I was diagnosed with ADHD'. For the control group, users are selected at random, and their public posts are inspected to ensure there are no posts of self-reported diagnosis of a behavioral disorder. The table 3-1 lists the total number of users, the average number of tweets per user, and the total number of tweets after preprocessing for the diagnosed group and the control group.

The data was collected between March 2017 to May 2017. For each user in the diagnosed group, a time T1 was set (as show in Figure 3-1). T1 indicates the date/time a user publicly states that they were or have been diagnosed with ADHD. Furthermore, the data before time T1 is referred to as a pre-diagnosed group, and the data after T1 is referred to as a post-diagnosed group.

Table 3-1: Data collection statistics for the diagnosed and control group.

	Diagnosed Group	Control Group
Total number of users	132	91
The average number of tweets per user	92	128
Total number of tweets	12,512	11,722

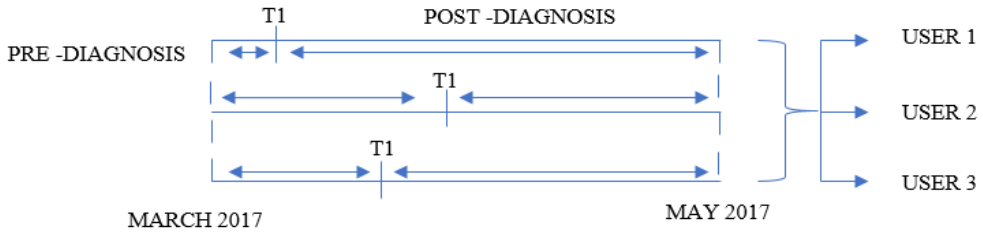

Figure 3-1: Diagrammatic representation of time T1.

3.2.3 Behavioral Measure 1: Variations in Phrase Structure Rules

Phrase structure rules are used to describe the syntax of a language and are closely associated with theoretical generative grammar. These rules can be categorical, rules that expand categories into other categories, or they can be lexical, rules that expand category labels by word.

The data transformation technique is replicated for the three groups of users, namely, the pre-diagnosed group, the post-diagnosed group, and the control group.

The tweets collected using the Twitter API are tokenized and categorized according to their part of speech tag using Noah's ARK by Carnegie Mellon. ARK uses the Penn Treebank tag set for categorizing tokens according to their parts of speech. The treebank consists of 33 parts of speech, including but not limited to adjectives, nouns, adverbs, and verbs. Each of these broad categories is represented by multiple tags denoting fine-grained

specifics of grammatical usage. For example, adjectives can be tagged as their base form or based on their intensity, comparative adjectives, or superlative adjectives.

The tagged tokens are stored in arrays, yielding 33 parts of speech arrays. Pairwise comparison of these arrays yields the cosine similarity between them. The cosine similarity values are stored in an NxN co-occurrence matrix, where N is the number of parts of speech tags.

A variation of the one hot matrix is used to obtain the absence or presence of a part of speech in a tweet. A 1 indicates the presence of a part of speech, and 0 indicates the absence of a part of speech in a tweet. This matrix is multiplied by the co-occurrence matrix. The resulting matrix is used as input to the recursive neural network. The recursive neural network used for this step is a stacked RNN with three layers, an embedding layer; a long short-term memory (LSTM) layer; and a dense layer. The LSTM layer has four components: a cell, an input gate; an output gate; and a forget gate. The three gates use a logistic function to compute an activation. The activation function is:

$$S(x) = \frac{1}{1 + e^{-x}} = \frac{e^{x}}{e^{x} + 1} \qquad \textbf{Eq 3.9}$$

The RNN runs for four epochs for the train and test set. The complexity of the above algorithm is O(n*h), where *m* is the number of hidden units, and *h* is the length of the epoch. The complexity of calculating the cosine similarity is $o(mn^2)$ where *m* is the number of terms that are common between two vectors, and n^2 is the number of iterations.

3.2.4 Behavioral Measure 2: Topic Detection

The tweets collected from Twitter for the three groups of users: pre-diagnosed, post-diagnosed, and the control group have been used for this step. The package provided by python, sci-kit learn, has been used to convert all the tweets into their respective TF-

IDF matrix; perform non-negative matrix factorization; and create, visualize the clusters using T-SNE. The methodology is repeated individually for the three groups of users.

The python package provides users with a TfidfVectorizer method to convert raw data into a matrix of tf-idf features. Two arguments are passed to the method, min_df and max_df, both of which are frequency parameters to be ignored if higher than or lower than the specified arguments. The resulting parameters are fit on the training set using fit_transform. Non-negative matrix factorization (NMF) is performed on the tf-idf matrix. The values for the arguments are set by experimentally determining the values. The arguments passed to NMF method are the number of components, random_state, solver, and beta_loss. Beta loss is passed to minimize the beta divergence, measuring the distance between the input matrix X and the dot product of WH. In this case, the number of components is set to 10, the solver is set to *mu*, random state is set to 7 and the beta loss is set to *kullback-leibler*. The result is fit to the training set using fit_transform and stored as W. The matrix H is set to the components of the result of NMF.

The top 10 words from each of the topics are chosen but since the result of NMF sorts the words in ascending order, the list must be first sorted in descending order. To visualize the clusters and to view the tweets in each cluster, TSNE and click events are used. The dimensions of the NMF matrix are reduced using TruncatedSVD. The number of components passed as an argument to the method is set to 50. The result is fit for the training set and TSNE is run on it. The scatter method provided by the matplotlib python package is used to show the clustered data points on a scatter plot. The argument s (represents the area) is set to 15 for the pre-diagnosed and post diagnosed group. For the control group, the argument s is set to 75.8.

3.2.5 Analysis: Sentiment and Emotion

The tokenized tweets tagged in the CONLL format are categorized into their own sentiments and emotion using the NRC word-emotion association lexicon. The lexicon identifies two sentiments: positive and negative. It also identifies eight emotions: anger, fear, anticipation, trust, surprise, sadness, joy, and disgust. The categorization of the tokens according to their emotions and sentiment, would aid in the creation of a timeline of the disorder for each user.

3.3 Results

3.3.1 Behavioral Measure 1: Variations in Phrase Structure Rules

The results of the recurrent neural network are reflective of the difference in the language used by users with the disorder and users without the disorder. Since parts of speech form the essential component of a sentence, the placement of a part of speech and its type in a sentence are essential to understand an individual's speech patterns. Figures 3-2 to 3-5 show the part of speech preferences for the three groups of users.

In the figures 3-2 to 3-5, the x-axis represents the parts of speech and the y-axis lists the results of the RNN (represents the importance of a part of speech as compared to the others). For the pre-diagnosed group, plural nouns (NNP) and prepositions (IN) are more likely to be used more frequently than a singular noun (NNS).

Similarly, for the control group, prepositions and modifiers (MD) are more likely to occur more frequently in a sentence than nouns (NN). In table 3-2, S represents an embedded clause.

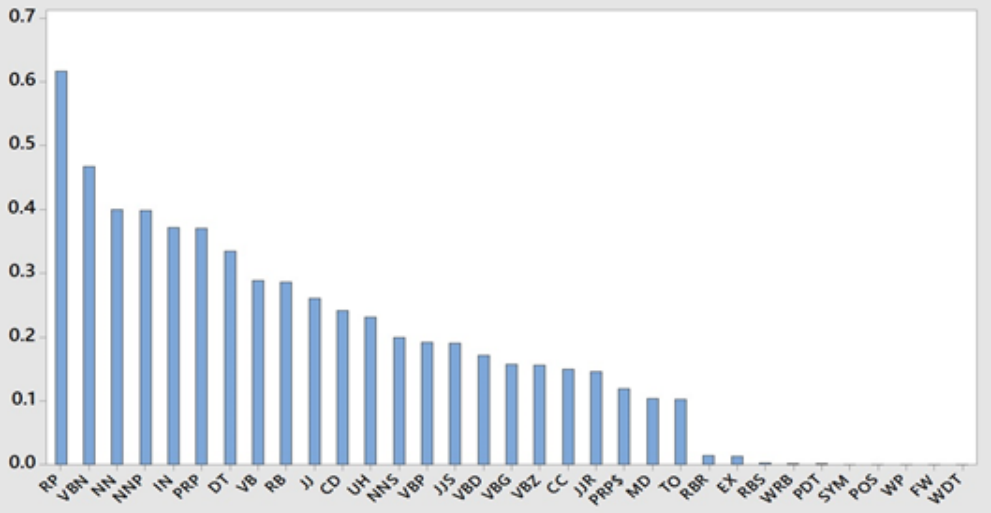

Figure 3-2: Results of RNN for pre-diagnosed group.

An embedded clause in a sentence is a group of words that include a subject and a verb, embedded within and dependent on the sentence's main clause. The pre-diagnosed group is more likely to use embedded clauses as compared to the other two groups, post-diagnosed and control group.

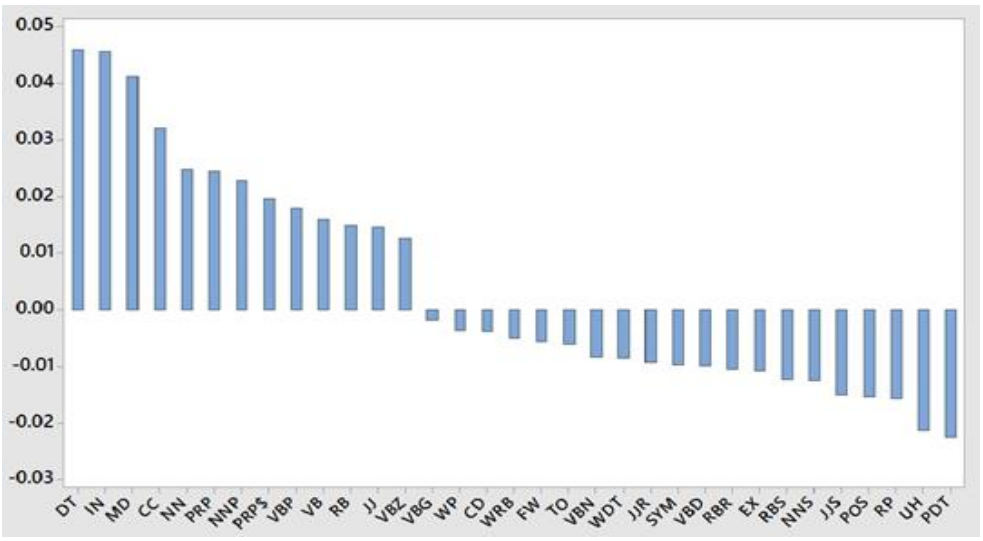

Figure 3-3: Results of RNN for post-diagnosed group.

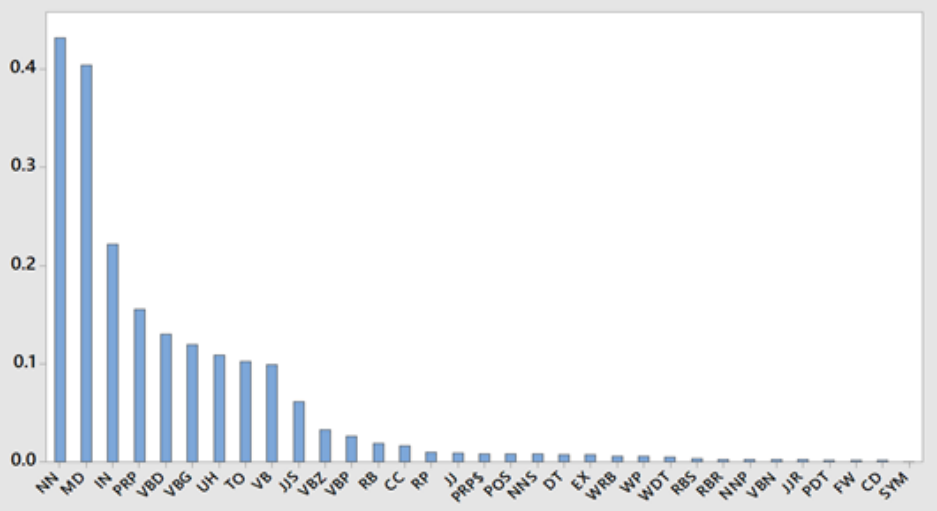

Figure 3-4: Results of RNN for control group.

Table 3-2: Formal rules for the English language.

Phrase Structure Rules
S→ NP (MD) VP
NP → V (NP) (AdjP) N (PP)
VP → V (NP) (PP) (AdvP)
AdvP → Adv (AdjP)
PP → P N
NP → N (Conj P)
VP → V (Conj P)
S → S (Conj S)

Furthermore, a logical combination of the parts of speech or group-specific phrase structure rules can elaborate on an individual user's speech pattern. The table 3-2, lists formal phrase structure rules for the English language. The usage of parts of speech in

brackets are optional. Parts of speech can be substituted in the phrase structure rule that defines it, or in another phrase structure rule.

The table 3-3 lists the phrase structure rules for the pre-diagnosed group, the post-diagnosed group, and the control group. To better understand the rules specific to the three groups of users, take, for example, the second rule from Table 3-2:

$$NP \rightarrow V\ (NP)\ (AdjP)\ N\ (PP)$$

As mentioned before, the parts of speech in parenthesis are optional. In the case of the pre-diagnosed group, NN (noun, basic form) and NNP (plural noun) are forms of the noun that are both frequently used, the rule becomes:

$$NP \rightarrow (NN/NNP)\ V\ (AdjP)\ (PP)$$

Prepositional phrases (PP) can be broken down into prepositions and nouns, according to the fifth rule in Table 3-6:

$$PP \rightarrow P\ N$$

Since, prepositions and nouns are important for the group, the rule now becomes:

$$NP \rightarrow (NN/NNP)\ V\ (AdjP)\ (PP)$$

Similarly, since verbs and adjectives aren't frequently used, the rule now becomes:

$$NP \rightarrow (NN/NNP)\ (PP)$$

For the control group, the parts of speech NN (noun, basic form) and NNS (singular noun) are the forms of the noun that are frequently used. Therefore, the rule becomes:

$$NP \rightarrow (NN/NNS)\ (PP)$$

The breakdown of the prepositional phrases (PP) remains the same in all cases.

Table 3-3: Phrase structure rules for the pre-diagnosed, post-diagnosed, and control group.

Pre-diagnosed Group	Post-diagnosed Group	Control Group
S → NP VP	S → NP VP	S → NP VP
NP → (NN/NNP) (PP)	NP → (NN/NNP) (PP)	NP → (NNS/NNP) (PP)
VP → V (RB)	VP → VB (RB) (NP)	VP → V (IN)
AdjP → Adj (AdvP)	AdjP → Adj (AdvP)	AdjP → Adj (AdvP)
AdvP → Adv (AdjP)		AdvP → Adv (AdjP)
PP → P N		PP → P N

3.3.2 Behavioral Measure 2: Topic Detection

The following are the clusters for the three groups of users: the green clusters are for the pre-diagnosed group; the blue clusters are for the post-diagnosed group, and the red clusters are for the control group. The clusters are based on the results of t-SNE. There are 128 pre-diagnosed clusters, 128 post-diagnosed clusters, and 72 control group clusters. Each data point in the cluster represents a tweet. The scatter plot (figure 3-5) has been obtained with the t-SNE perplexity set to 10. Since there is no natural separation of the data, a supervised approach in chapter 4 is the best approach in finding the differences between the three groups of users.

Since Twitter is an informal platform, a separation in the users based on the topics is hard to achieve. On a finer level, such as parts of speech or sentiment, separation is possible and has been shown in this chapter.

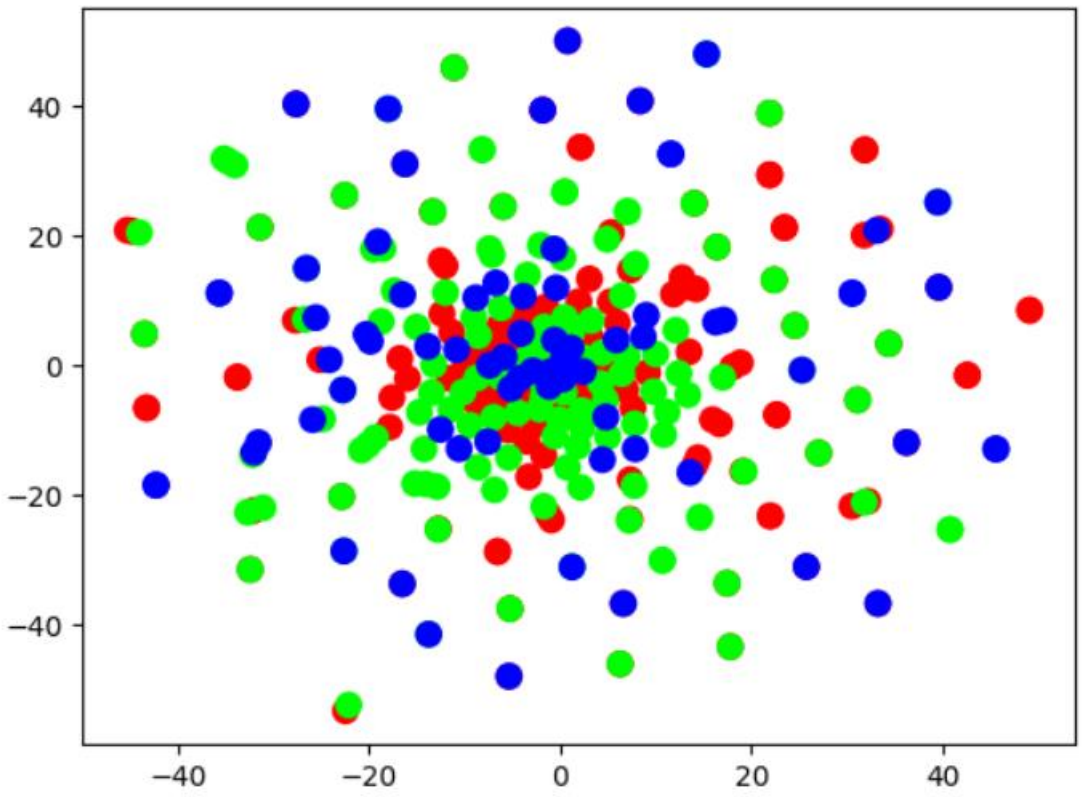

Figure 3-5: Scatter plot for topic detection.

The following (tables 3-4 to 3-6) are the top ten topics for the three groups of users.

Table 3-4: Top ten topics for the pre-diagnosed group.

Topic	Pre-diagnosed Group
1	all,at,an,because,we,ll,could,or,up,out
2	cbd,via,anyone,best,help,new,from,high,or,does
3	what,do,know,fuck,even,did,say,about,dont,name
4	oh,did,shirotwt,fucking,as,thank,right,yeah,well,lol
5	one,he,been,no,by,has,got,had,out,now

Table 3-4: Top ten topics for the pre-diagnosed group.

Topic	Pre-diagnosed Group
6	about,how,know,really,feel,think,no,much,make,too
7	love,im,too,back,go,much,ok,its,koutameoshi,your
8	get,good,at,do,also,they,need,some,look,anewrecipeh
9	don,am,got,need,want,adhd,here,any,game,some
10	do,when,kingwilliamiv3,typicalgamer,samararedway,couldnotagree,as,your,lol,them

One of the top ten topics for the pre-diagnosed group consists of items related to drugs. This is indicative of their inclination towards using drugs. Both the pre-diagnosed and post-diagnosed groups of users have items related to ADHD as one of their top ten topics (#8 for the pre-diagnosed group in table 3-4 and #2 for the post-diagnosed group in table 3-5).

Table 3-5: Top ten topics for the post-diagnosed group.

Topic	Post-diagnosed Group
1	at,we,by,from,as,amp,lol,need,make,out
2	im,its,gonna,back,dante,see,off,hate,life,cant
3	good,an,day,adhd,aspie,again,night,myself,am,got
4	love,more,much,too,avi_kaplan,as,great,happy,beautiful,amazing
5	don,want,anyone,they,even,because,does,or,about,re
6	up,he,fuck,from,look,cute,shit,nice,him,dante
7	get,who,his,he,armyofkek,buy,him,wants,please,come
8	how,no,am,now,oh,thank,right,feel,god,graysondolan

Table 3-5: Top ten topics for the post-diagnosed group**.**

Topic	Post-diagnosed Group
9	about,at,time,been,one,has,ve,first,best,found
10	when,know,fucking,had,people,dont,man,ass,out,looks

One of the top ten topics for the pre-diagnosed group consists of items related to drugs. This is indicative of their inclination towards using drugs. Both the pre-diagnosed and post-diagnosed groups of users have items related to ADHD as one of their top ten topics (#8 for the pre-diagnosed group in table 3-4 and #2 for the post-diagnosed group in table 3-5).

Table 3-6: Top ten topics for the control group.

Topic	**Post-diagnosed Group**
1	at,we,by,from,as,amp,lol,need,make,out
2	im,its,gonna,back,dante,see,off,hate,life,cant
3	good,an,day,adhd,aspie,again,night,myself,am,got
4	love,more,much,too,avi_kaplan,as,great,happy,beautiful,amazing
5	don,want,anyone,they,even,because,does,or,about,re
6	up,he,fuck,from,look,cute,shit,nice,him,dante
7	get,who,his,he,armyofkek,buy,him,wants,please,come
8	how,no,am,now,oh,thank,right,feel,god,graysondolan
9	about,at,time,been,one,has,ve,first,best,found
10	when,know,fucking,had,people,dont,man,ass,out,looks

In contrast, the top ten topic lists for the control group contains items probably related to topics that are or were the news. The emotions expressed, whether negative or positive, are concerning the news items.

3.3.3 Analysis: Sentiment and Emotion

The result set for the pre-diagnosed users in figure 3-6 shows that their expressed sentiments are more negative. Similarly, the emotions most commonly expressed by pre-diagnosed users are anger and sadness.

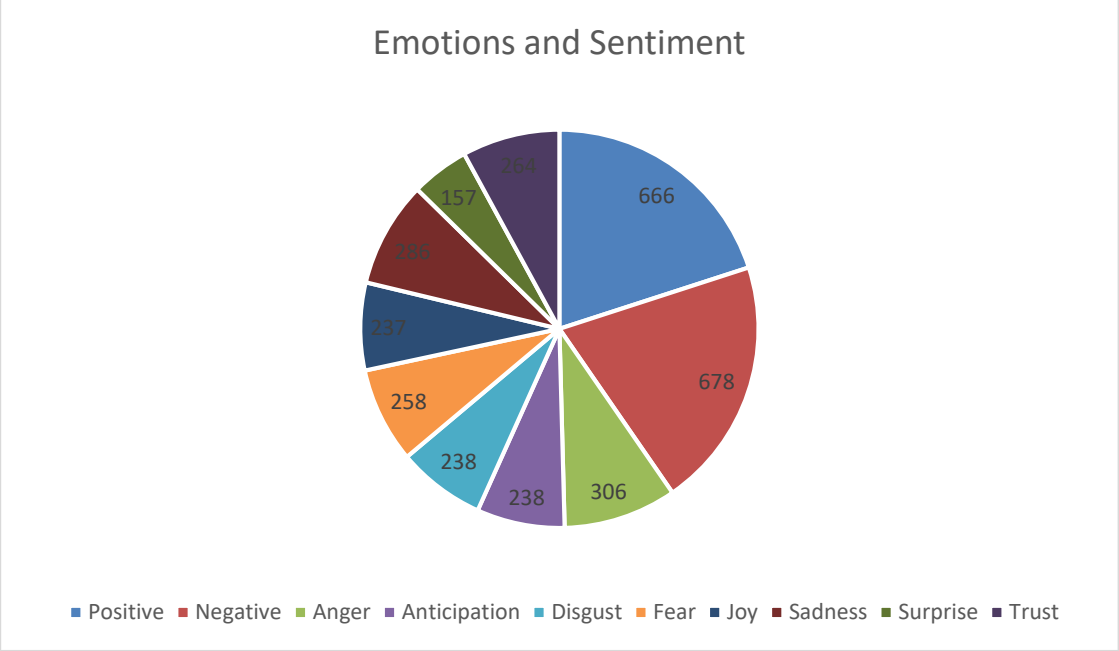

Figure 3-6: Emotion and sentiment for the pre-diagnosed group.

The result set for the post-diagnosed group of users in figure 3-7 shows that the sentiment for their posts is more positive, and the emotions expressed are trust and fear.

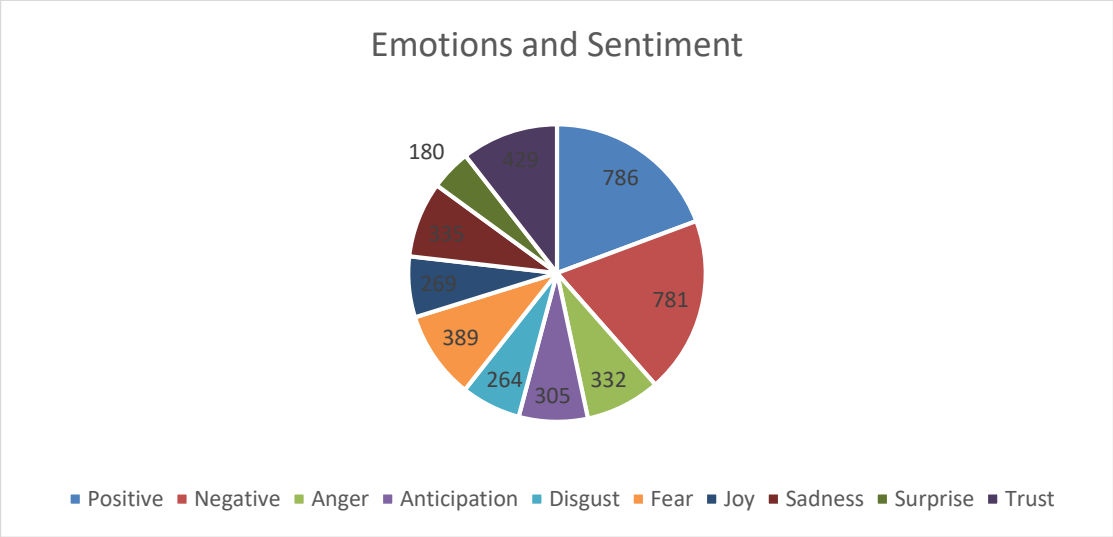

Figure 3-7: Emotion and sentiment for the post-diagnosed group.

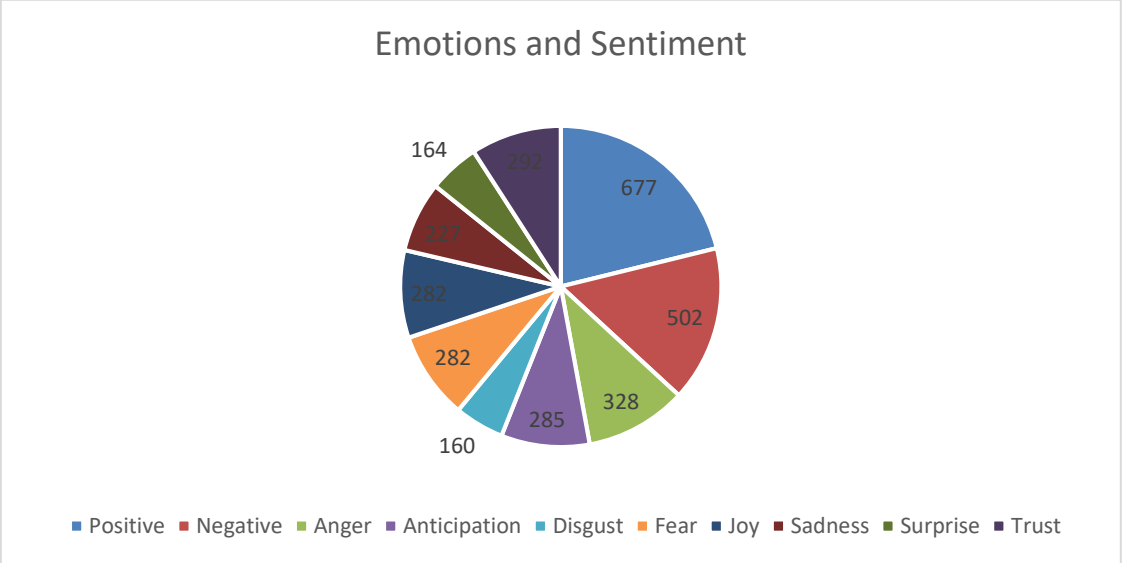

Figure 3-8: Emotion and sentiment for the control group.

The result set for the control group in figure 3-8 shows that users in the group tend to be more positive in their outlook, and the emotions most commonly expressed by them are anger and trust.

3.4 Conclusion

In conclusion, the two measures and the analysis show a stark contrast in three groups of users, pre-diagnosed, post-diagnosed, and control group. These measures and analyses can be further exploited to identify critical points in the timeline of the disorder: the peak and time to remission.

CHAPTER 4

CLINICAL DECISION SUPPORT SYSTEM FOR ADHD

The United States alone reports three million cases of ADHD every year (Data and Statistics about ADHD, 2019). The commonplace nature of the disorder and the longstanding societal stigma associated with it leaves many more cases undiagnosed. The ability to use social media to identify users with ADHD can assist clinicians in diagnosing patients in remote areas or areas with a deep understanding of the disorder. It has the potential to improve the specificity and sensitivity of ADHD detection. An effective clinical decision support system can allow monitoring of patient's adherence to prescribed treatment options. It can also establish a hypothesis for future clinical and research investigations in the future.

This chapter focuses on a clinical decision support system for the disorder using a classification algorithm, decision trees. The chapter is organized as follows: related works, definitions, equations and algorithms, methodology, results, and conclusion.

4.1 Related Works

Recently, there has been an increase in research using language to identify people with mental illnesses and quantify its progression. De Choudhury, et al. (2013b) worked on identifying and helping people who suffered from depression. Cloninger, et al. (1993) evaluated the personality traits that made people vulnerable to depression. Rude, et al. (2003) successfully hypothesized that negative processing biases in resolving verbal cues

could predict future episodes of depression. Brown, et al. (1990) found that the lack of support from peers and low self-esteem leads to higher incidences of depression. Paul and Dredze were able to learn more about diseases from posts obtained from Twitter (Paul & Dredze, 2011). Kotikalapudi, et al. (2012) hypothesized that an analysis of web activity of college students could identify users with depression. Moreno, et al. (2011) proved that updates on Facebook could reveal symptoms of depression.

Coppersmith, et al. (2014) researched methods to identify people with post-traumatic stress disorder. Disease surveillance on social media was explored by Brownstein, et al. (2009). The ample data available on social media was explored by Paul and Dredze (2011).

4.2 Methodology

The analysis for this chapter is on a user-level and a tweet-level. The data used for this analysis is the same as the Twitter user posts used in Chapters 2 and 3. There are four main categories for the features used for this analysis: TF-IDF; topic detection clusters, parts of speech, and sentiment, and emotion. The methodology to obtain the features is repeated for the three groups of users: pre-diagnosed, post-diagnosed, and the control group. A decision tree is used to predict classes (pre-diagnosed, post-diagnosed, and the control group) for the test set.

4.2.1 Definitions, Equations, and Algorithms

Definition 4.2.1 Decision Tree is a tree-like model of decisions and possible consequences, commonly used in decision analysis and operations research. A decision tree consists of three types of nodes: decision nodes, chance nodes, and end nodes (Witten, et al., 2016).

Definition 4.2.2 Entropy is a measure of the disorder (Witten, et al., 2016). The formula for entropy is:

$$E(S) = \sum_{i=1}^{c} -p_i \log_2 p_i \qquad \textbf{Eq 4.1}$$

where p_i is the frequentist probability of a class in the data set.

Definition 4.2.3 F1 Score is a measure of the accuracy of a test (Witten, et al., 2016). It considers precision (p) and recall (r) of the test to compute the score. The traditional formula for the test is:

$$F_1 = \left(\frac{2}{recall^{-1} + precision^{-1}}\right) = 2.\frac{precision.recall}{precision + recall} \qquad \textbf{Eq 4.2}$$

4.2.2 TF-IDF

The statistic TF-IDF is calculated using the sci-kit learn package in Python. For the users and tweets, their tweets are fed as input to the Tfidfvectorizer method provided by the package. The arguments for the method, min_df, and max_df are set to 0.01 and 1, respectively. The argument binary is set to true. The initial parameters are fit on the data set and transformed using the fit_transform method provided by the same package. Since the resulting matrix is a sparse matrix, the todense method is used to return a dense representation of the matrix. The final array is collapsed into a 1D array. This results in a 132x14027 matrix for the pre-diagnosed group, a 132x16384 matrix for the post-diagnosed group, and a 91x2400 matrix for the control group.

4.2.3 Topic Detection Clusters

The tfidf matrix is used as input to the non-negative matrix factorization algorithm, provided by the NMF package provided by Scikit-learn. The arguments for the methods are assigned experimentally validated values. The arguments for the NMF method,

random_state, beta_loss, solver are set to 7, kullback-leibler and mu, respectively. The initial parameters are fit on the data set and transformed using the fit_transform package provided by the package, resulting in the matrix W. The components from the result are saved as matrix H. The indices of the maximum values along axis 0 in matrix X are saved and used for the decision tree. This results in a 132x1350 matrix for the pre-diagnosed group, a 132x16384 matrix for the post-diagnosed group, and a 91x6753 matrix for the control group.

4.2.4 Parts of Speech

The parts of speech categories used for this step are from the Penn Tree Bank. Thirty-three parts of speech are considered. To categorize tokens according to their parts of speech, the NOAH's ARK by Carnegie Mellon is used. For each of the users in the three groups, the total count for each part of speech is calculated. This results in a 132x32 matrix for the pre-diagnosed and post-diagnosed group and 91x32 matrix for the control group.

4.2.5 Sentiment and Emotion

The NRC emotion lexicon is used to categorize user tweets into two sentiments and eight emotions. The syuzhet package in R is used to categorize the tweets. This results in a 132x10 matrix for the pre-diagnosed and post-diagnosed group, and a 91x10 matrix for the control group.

Table 4-1: Algorithm for calculating sentiment and emotion.

Algorithm: Sentiment and emotion for pre-diagnosed, post-diagnosed and the control group.
Input: User tweets for pre-diagnosed, post-diagnosed, and control group.

Output: Total count of sentiment and emotion for each user in the pre-diagnosed, post-diagnosed and control group.

1. **FOR** i in list of user files

    Read *data* for file i

    nrc_data = get_nrc_sentiment(data)

    Save/Append data to csv file.

    ENDFOR

    End

4.2.6 Decision Tree

The decision tree is implemented in Python using the Scikit-learn package. The decision tree algorithm uses the GINI index to build the tree:

$$Gini = 1 - \sum_{i=1} (p_i)^2$$

where p_i denotes the probability of the classes.

The input to the decision tree is an aggregation of the matrices for the four features, tfidf, topic detection clusters, parts of speech, and sentiment and emotion. There are three classes for this classification algorithm, pre-diagnosed (0), post-diagnosed (1), and the control group (2). The first column for the matrix for the pre-diagnosed group is set to 0, the post-diagnosed group is set to 1, and the control group is set to 2. Four cases are considered, pre-diagnosed vs. post-diagnosed, pre-diagnosed vs. control, post-diagnosed vs. control, pre-diagnosed vs. post-diagnosed vs. control.

Table 4-2: Decision tree classifier for pre-diagnosed, post-diagnosed, and control Group.

Algorithm: Decision tree classifier for pre-diagnosed, post-diagnosed and control group.

Input: Matrices for the user groups (pre-diagnosed + post-diagnosed, pre-diagnosed + control + post-diagnosed + control, pre-diagnosed + post-diagnosed + control).

Output: Predicted values for classifying data by group.

1. Concatenate matrices for user groups and save as matrix X.
2. Shape matrices into 2D arrays and save as matrix Y.
3. Split X and Y into train and test set using a 70/30 split. Set a random state to 100.
4. Run the decision classifier with argument criterion = “entropy”, random_state = 10, max_depth = 3, min_samples_leaf = 5. Save as cli_entropy.
5. Fit X_train, Y_train on cli_entropy.
6. Predict class values for X_test.
7. **End**

For the first three cases, the matrices are loaded two at a time. The matrices are concatenated (matrix X), and an array is created from X (matrix Y). The matrices X and Y are randomly split into train and test sets, X_train, Y_train, X_test, and Y_test. This is accomplished using the train_test_split method provided by the Scikit-Learn package. The arguments for the method, test_size, random_state are set to 0.3 and 100, respectively.

The train test size method splits the dataset into 70% train and 30% test. The random state argument randomly selects the values to split into train and test sets. The decision tree classifier method provided by the same package is used to construct the decision tree. The arguments for the method criterion, random_state, max_depth, and min_samples_leaf are set to entropy, 10, 3 and 5, respectively. The final matrix is obtained by fitting the X_train and Y_train matrix on the result of the decision tree classifier. The predictions for X_test can be obtained by using the predict method.

4.3 User-Level Results

4.3.1 Pre-diagnosed group vs. Post-diagnosed group

The decision tree in figure 4-1 lists the class separation for the pre-diagnosed group and the post-diagnosed group. The dependent variable of the decision tree (the root) has 140 observations and two classes, true or false. Entropy is the measure of impurity, disorder, or uncertainty in the samples. It controls how the decision tree splits the data. The highest feature is feature 18243 in the dataset. This corresponds to the emotion '*joy*'. Similarly, the feature 18236 is '*üzülmedim*,' the feature 18253 is '*foreign word*', the feature 18260 is '*singular noun*', the feature 18246 is '*trust*', the feature 18239 is '*anger*'. The F1 of the decision tree is 80%.

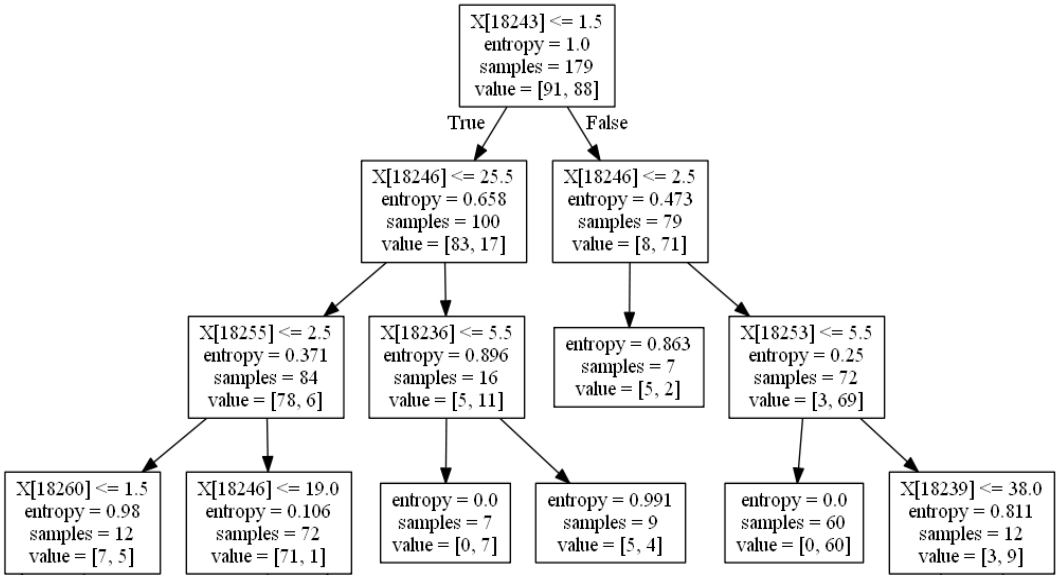

Figure 4-1: Decision tree for pre-diagnosed vs post-diagnosed group.

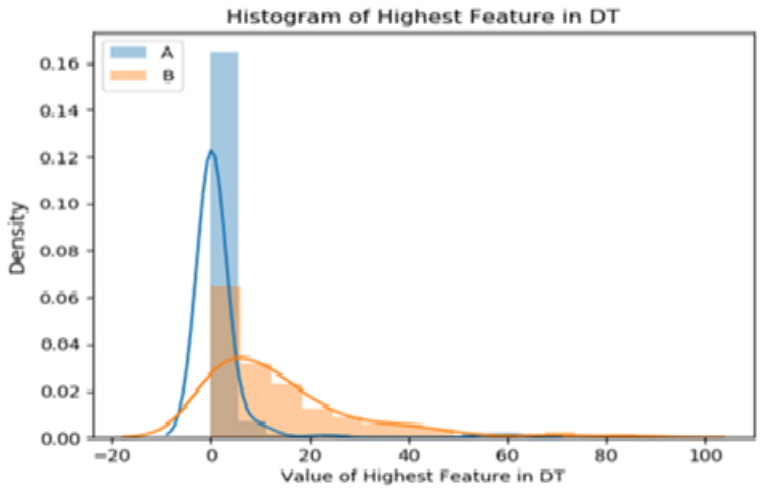

Figure 4-2: Histogram of the highest feature for pre-diagnosed vs. post-diagnosed group.

In figure 4-2, A represents the pre-diagnosed group, and B represents the post-diagnosed group. The histogram for the pre-diagnosed vs. post-diagnosed group is skewed right and therefore is asymmetrical. The graph shows a higher number of occurrences for the highest feature in the pre-diagnosed group.

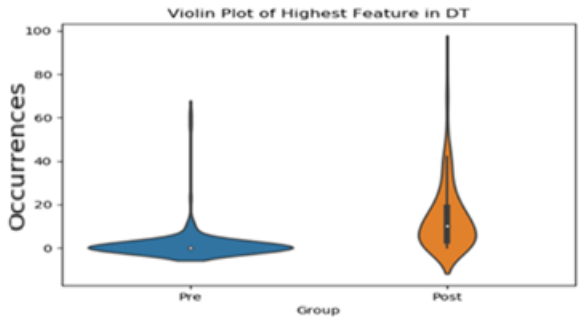

Figure 4-3: Violin plot of highest feature for pre-diagnosed vs. post-diagnosed group.

In figure 4-3, the violin plot shows the value of the highest feature. The median value (represented by the white dot in the middle) for the pre-diagnosed group is 0, and the post-diagnosed group is 10. The black bar is the interquartile range. The broader sections of the plot represent the occurrences of the highest feature in the pre-diagnosed group and the post-diagnosed group. From the above graph, almost all the occurrences of pre-diagnosed users are concentrated around the median.

4.3.2 Pre-diagnosed group vs. Control Group

The decision tree in figure 4-4 shows the class separation between the pre-diagnosed group and the control group. The dependent variable of the decision tree (the root) has 140 observations and two classes, true or false. The highest feature is feature 11505 in the dataset. This corresponds to the part of the speech '*singular noun*'. Similarly, feature 11502 is '*superlative adjective*', feature 11495 is '*determiner*,' feature 11507 is '*predeterminer*', feature 11481 is '*üzerinden'*, feature 11506 is 'plural noun'. The F1 score of the decision tree is 79%.

The standard label in figure 4-5 represents the control group. The histogram for the pre-diagnosed vs. control group is skewed right and therefore is asymmetrical. The graph shows a higher number of occurrences for the highest feature in the control group.

In figure 4-6, the standard label in the violin plot represents the control group. The above violin plot shows the value of the highest feature. The median value (represented by the white dot in the middle) for the control group is 0, and the pre-diagnosed group is in the range 0 to 100. The black bar is the interquartile range. The broader sections of the plot represent the occurrences of the highest feature in the pre-diagnosed group and the post-diagnosed group. From the above graph, almost all the occurrences of control group users are concentrated around the median.

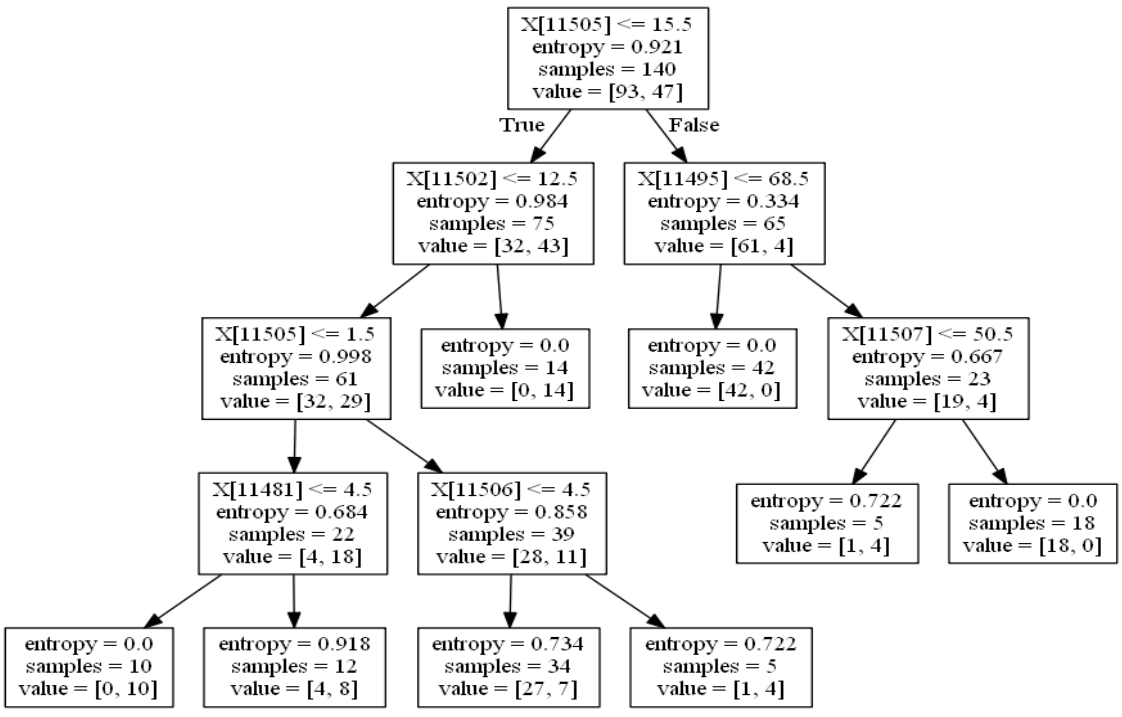

Figure 4-4: Decision tree for pre-diagnosed group vs. control group.

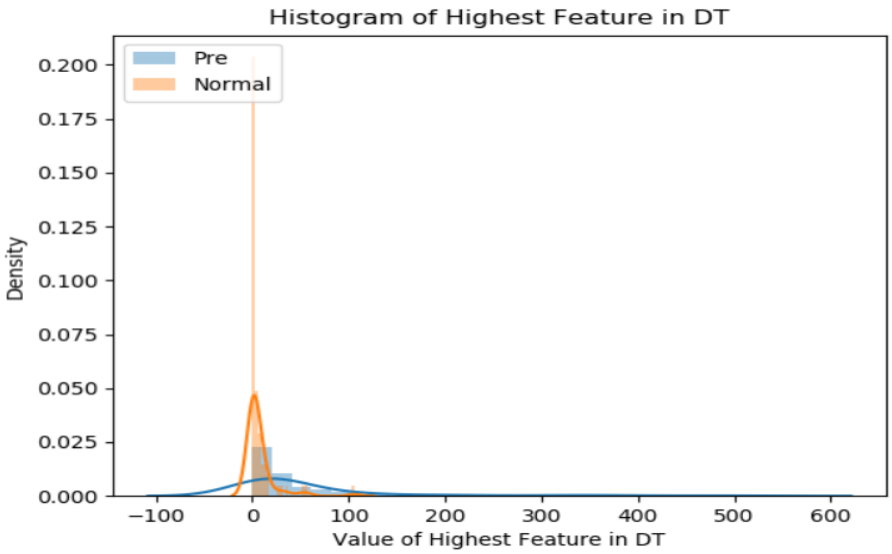

Figure 4-5: Histogram of the highest feature for pre-diagnosed group vs. control group.

The standard label in figure 4-5 represents the control group. The histogram for the pre-diagnosed vs. control group is skewed right and therefore is asymmetrical. The graph shows a higher number of occurrences for the highest feature in the control group.

In figure 4-6, the standard label in the violin plot represents the control group. The above violin plot shows the value of the highest feature. The median value (represented by the white dot in the middle) for the control group is 0, and the pre-diagnosed group is in the range 0 to 100. The black bar is the interquartile range. The broader sections of the plot represent the occurrences of the highest feature in the pre-diagnosed group and the post-diagnosed group. From the above graph, almost all the occurrences of control group users are concentrated around the median.

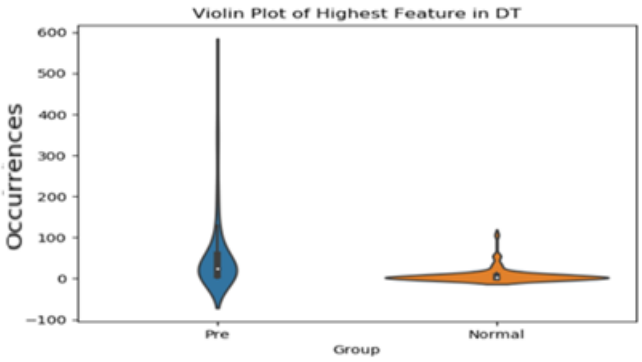

Figure 4-6: Violin plot of highest feature for pre-diagnosed group vs control group.

4.3.3 Post-diagnosed group vs. Control Group

The decision tree in figure 4-7 shows the class separation between the post-diagnosed group and the control group. The dependent variable of the decision tree (the root) has 140 observations and two classes, true or false. The highest feature is feature 15477 in the dataset. This feature corresponds to the emotion '*sadness*'. The feature 15493 corresponds to the part of speech '*singular noun*'. Similarly, the feature 15494 is '*plural noun*', the feature 15479 is '*trust*', the feature 15503 is '*superlative adverb'*. The F1 Score for the decision tree is 93%.

In figure 4-8, the normal label in the histogram represents the control group. The histogram for the pre-diagnosed vs. control group is skewed right and therefore is asymmetrical. The graph shows a density for the highest feature in the control group.

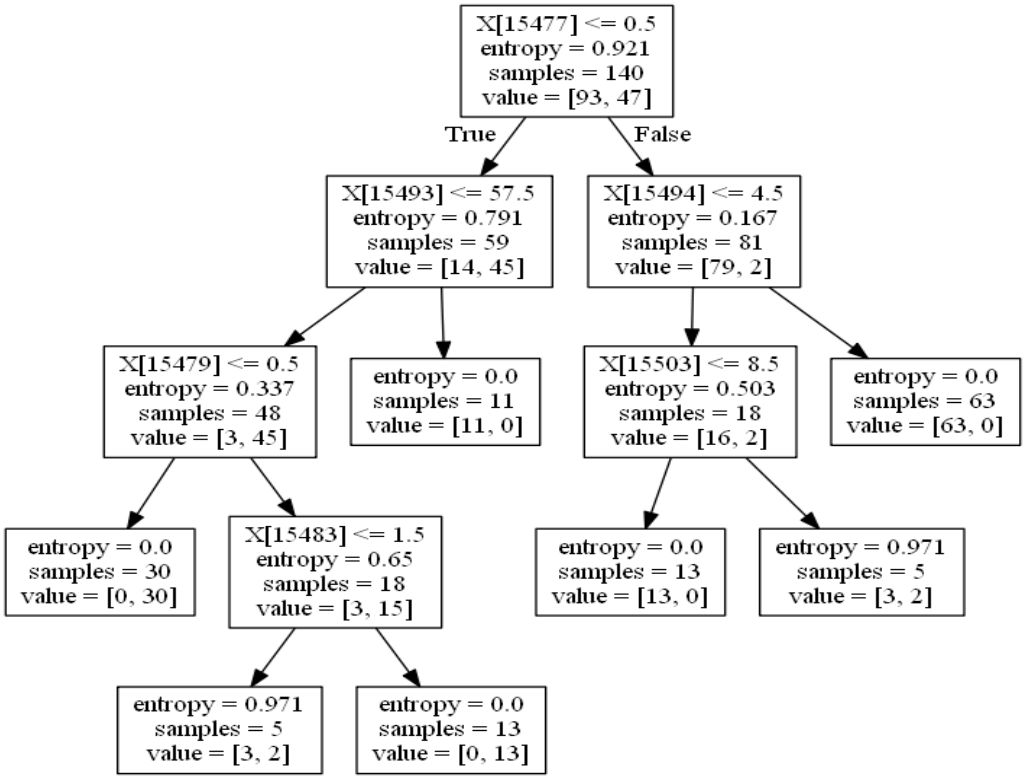

Figure 4-7: Decision tree for post-diagnosed group vs. control group.

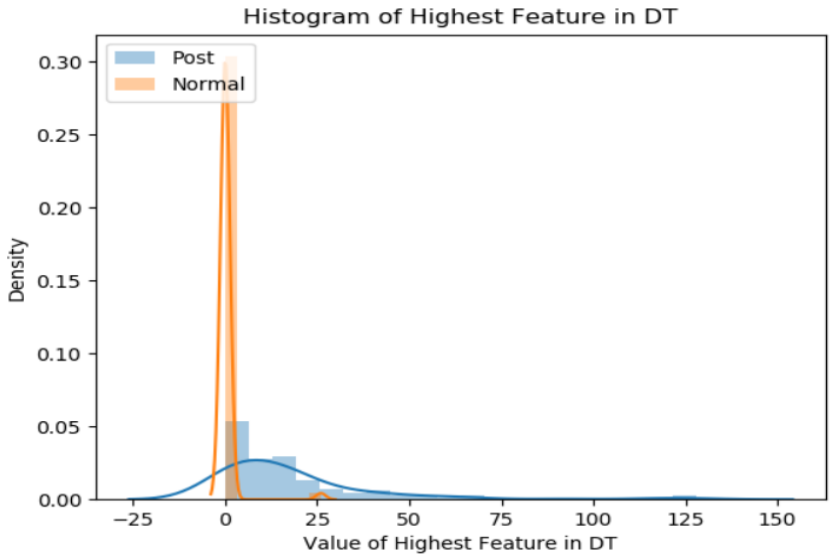

Figure 4-8: Histogram for the highest feature for post-diagnosed group vs control group.

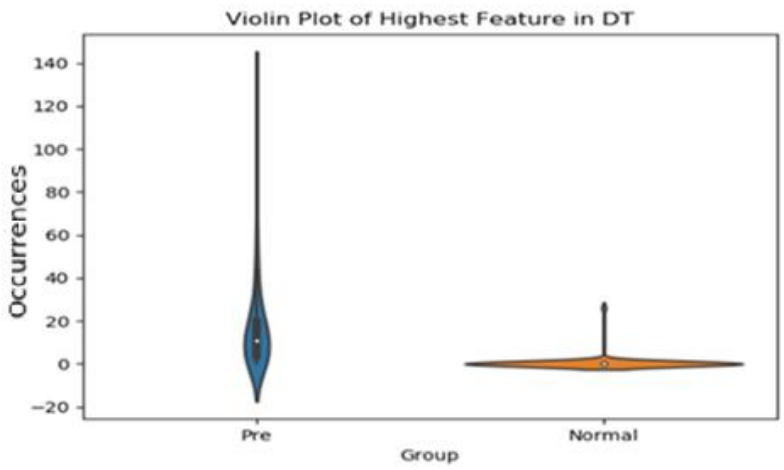

Figure 4-9: Violin plot for the highest feature for post-diagnosed group vs. control group.

In figure 4-9, the normal label in the violin plot represents the control group. The above violin plot shows the value of the highest feature. The median value (represented by the white dot in the middle) for the control group is 0, and the post-diagnosed group is in the range 0 to 20. The black bar is the interquartile range. The broader sections of the plot represent the occurrences of the highest feature in the pre-diagnosed group and the post-diagnosed group. From the above graph, almost all the occurrences of control group users are concentrated around the median.

4.3.4 Pre-diagnosed group vs Post-diagnosed group vs Control Group

The decision tree in figure 4-10 shows the class separation between the pre-diagnosed group, post-diagnosed group, and the control group. The dependent variable of the decision tree (the root) has 229 observations and two classes, true or false. The highest feature is feature 15478 in the dataset. This corresponds to the word '*seaham*' in the tfidf matrix. Similarly, the feature 15493 is '*sebbdavies*', the feature 15479 is '*seahorses*', the feature 15477 is '*seagal*', the feature 15506 is '*seda_ozen*', the feature 15490 is '*seatbelt*',

the feature 15481 is '*sedativeboy*', the feature 15490 is '*sealed*'. The F1 Score of the decision tree is 70%.

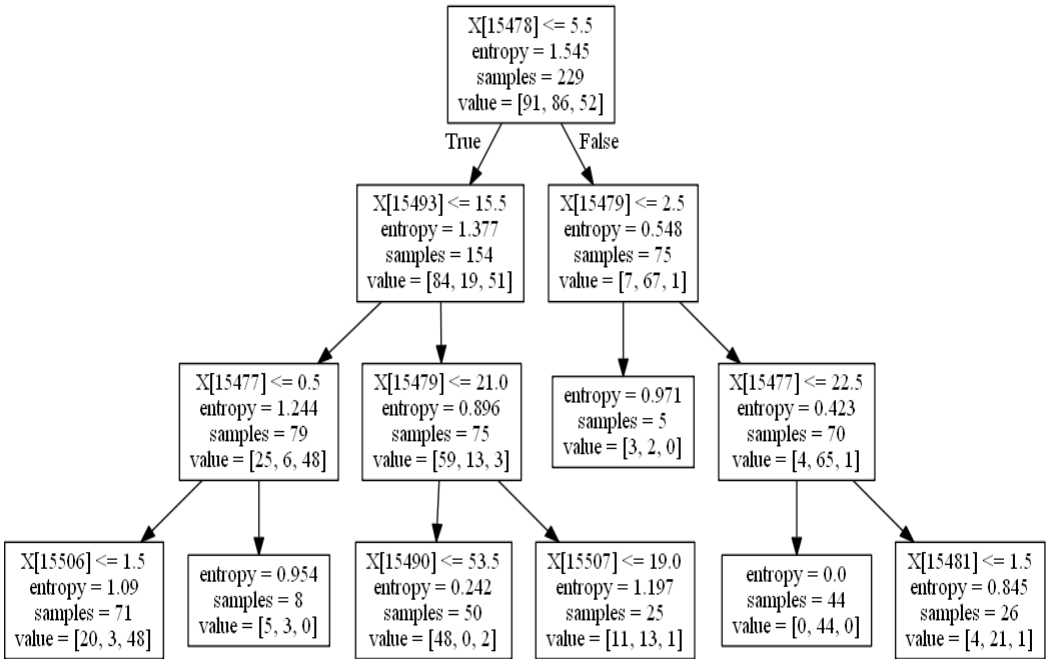

Figure 4-10: Decision tree for pre-diagnosed group vs. post-diagnosed group vs control group.

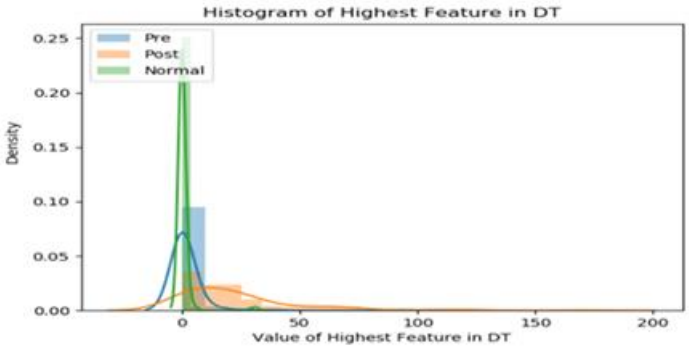

Figure 4-11: Histogram of highest feature for pre-diagnosed group vs post-diagnosed group vs control group.

In the figure 4-11, the normal label in the histogram represents the control group. The histogram for the pre-diagnosed vs. control group is skewed right and therefore is asymmetrical. The graph shows a higher density for the highest feature in the control group.

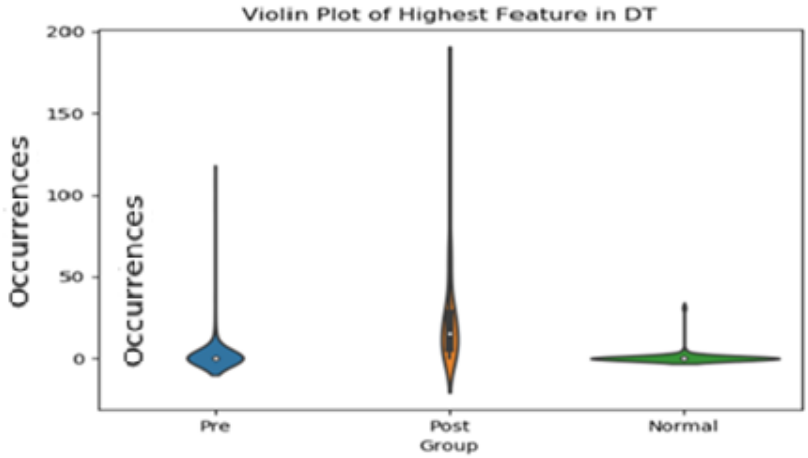

Figure 4-12: Violin plot of highest feature for pre-diagnosed group vs. post-diagnosed group vs. control group.

In figure 4-12, the normal label in the violin plot represents the control group. The above violin plot shows the values of the highest feature. The median value (represented by the white dot in the middle) for the control group and the pre-diagnosed group is 0, and the post-diagnosed group is in the range 0 to 50. The black bar is the interquartile range. The broader sections of the plot represent the occurrences of the highest feature in the pre-diagnosed group and the post-diagnosed group. From the above graph, almost all the occurrences of the control group and pre-diagnosed group users are concentrated around the median.

4.4 Tweet-Level Results

4.4.1 Pre-diagnosed group vs. Post-diagnosed group

The decision tree in figure 4-13 shows the tweet-level class separation between the pre-diagnosed group and the post-diagnosed group. The dependent variable of the decision tree (the root) has 17624 observations and two classes, true or false. The highest feature is feature 22683 in the dataset. This corresponds to the part of the speech '*adjective*'. Similarly, the feature 22688 is '*singular noun*', the feature 22685 is '*superlative adjective*', feature 22672 is '*sadness*', the feature 22670 is '*trust*', feature 22684 is '*comparative adjective*', feature 22702 is '*verb*', feature 22701 is '*interjection*', feature 22689 is '*plural noun*', feature 22690 is '*predeterminer*', feature 22693 is '*personal pronoun*', feature 22673 is '*joy*'. The F1 Score of the decision tree is 76%.

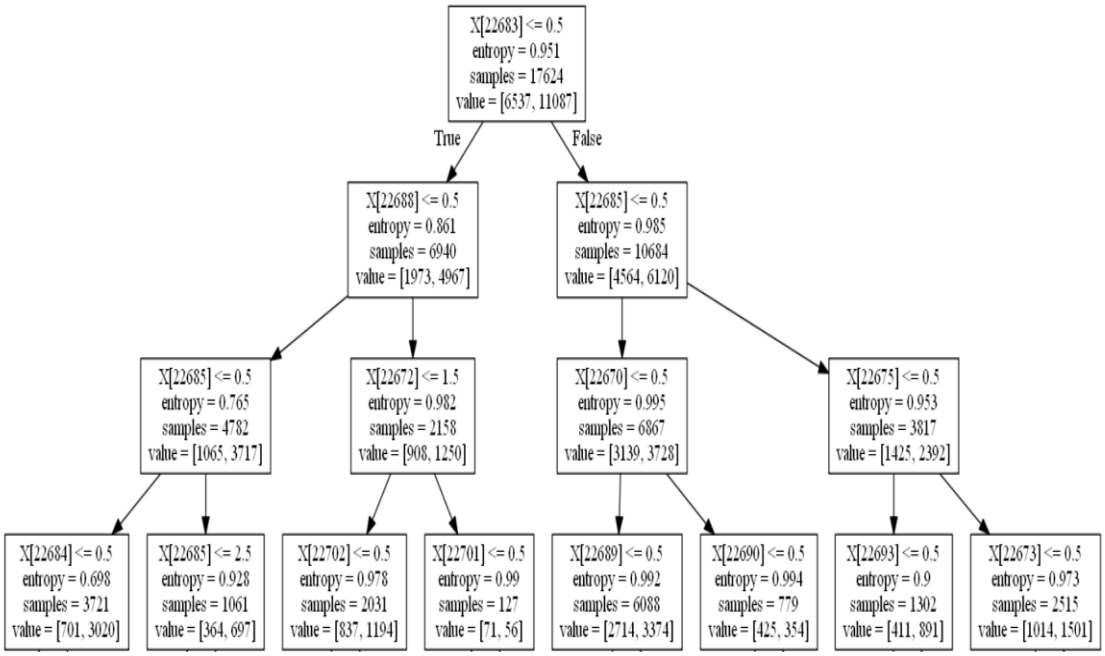

Figure 4-13: Tweet-level decision tree for pre-diagnosed group vs post-diagnosed group.

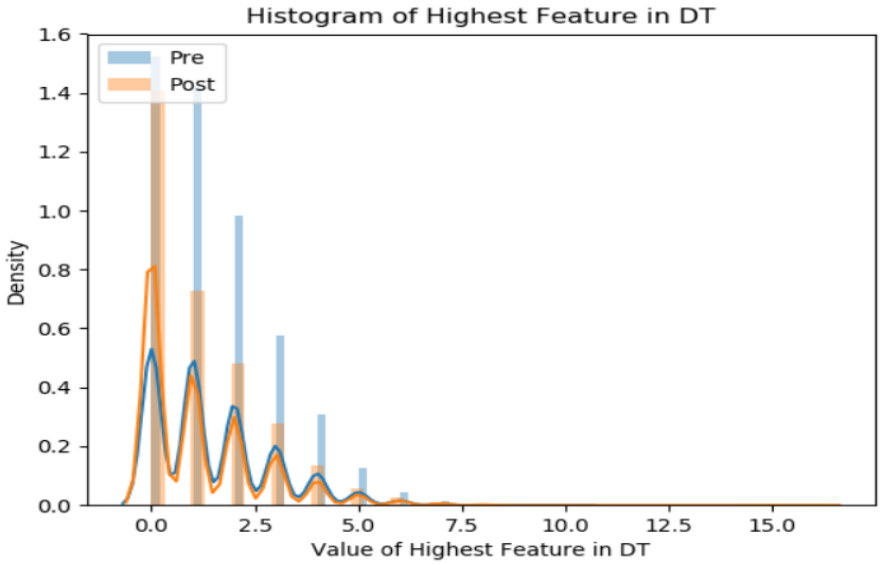

Figure 4-15: Tweet-level histogram of highest feature for pre-diagnosed group vs. post-diagnosed group.

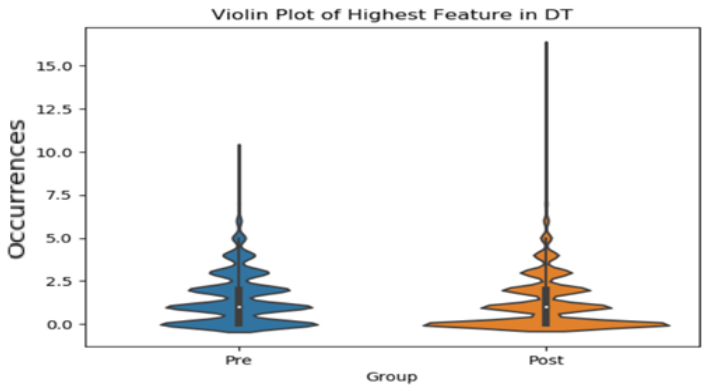

Figure 4-14: Tweet-level violin plot of highest feature for pre-diagnosed group vs post-diagnosed group.

In figure 4-14, the histogram for the pre-diagnosed vs. post-diagnosed group on a tweet level is skewed right and therefore is asymmetrical. The graph shows an equivalent

density for the pre-diagnosed and post-diagnosed group if the feature value is 0. For feature values higher than 0, the pre-diagnosed group has a higher density.

In figure 4-15, the violin plot shows the density of data at different values. The median value (represented by the white dot in the middle) for the control group and the pre-diagnosed and post-diagnosed group is greater than 0. The black bar is the interquartile range. The broader sections of the plot represent the occurrences of the highest feature in the pre-diagnosed group and the post-diagnosed group.

4.4.2 Pre-diagnosed group vs Control group

The decision tree in figure 4-16 shows the tweet-level class separation between the pre-diagnosed group and the control group. The dependent variable of the decision tree (the root) has 10903 observations and two classes, true or false. The highest feature is feature 19666 in the dataset. This corresponds to the part of the speech '*superlative adjective*'. Similarly, the feature 19669 is '*singular noun*', feature 19655 is '*disgust*. The F1 Score of the decision tree is 72%.

In figure 4-17, the label normal in the histogram represents the control group. The histogram for the pre-diagnosed vs. control group on a tweet level is skewed right and therefore is asymmetrical. The graph shows a higher density for pre-diagnosed if the feature value is 0. The density for the pre-diagnosed and control group is almost the same for feature values close to 1.25. For feature values higher than 1.25, the control group has a higher feature value.

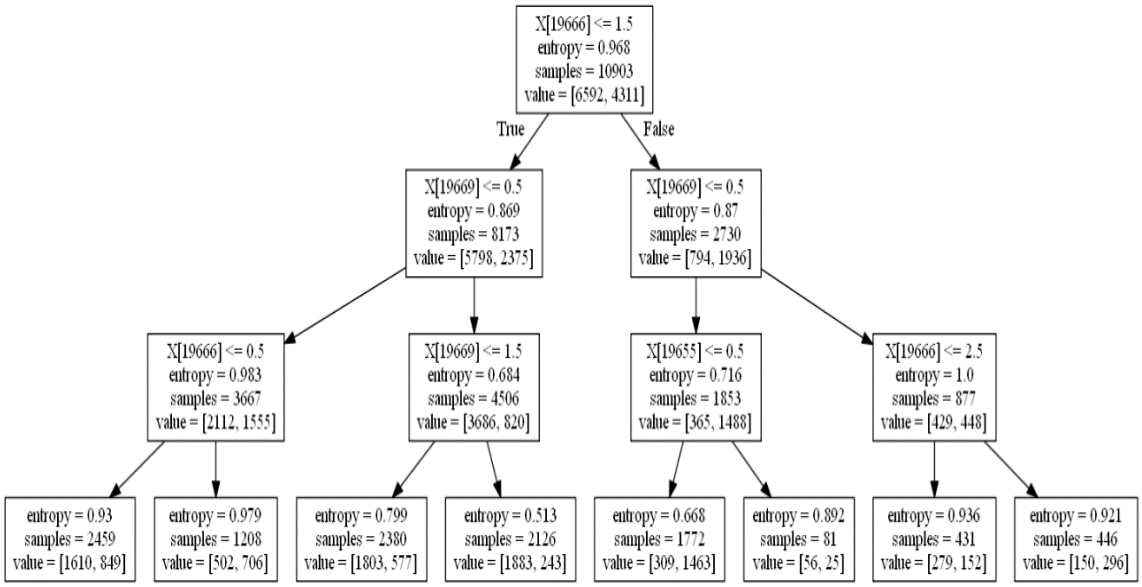

Figure 4-16: Tweet-level decision tree for pre-diagnosed group vs control group.

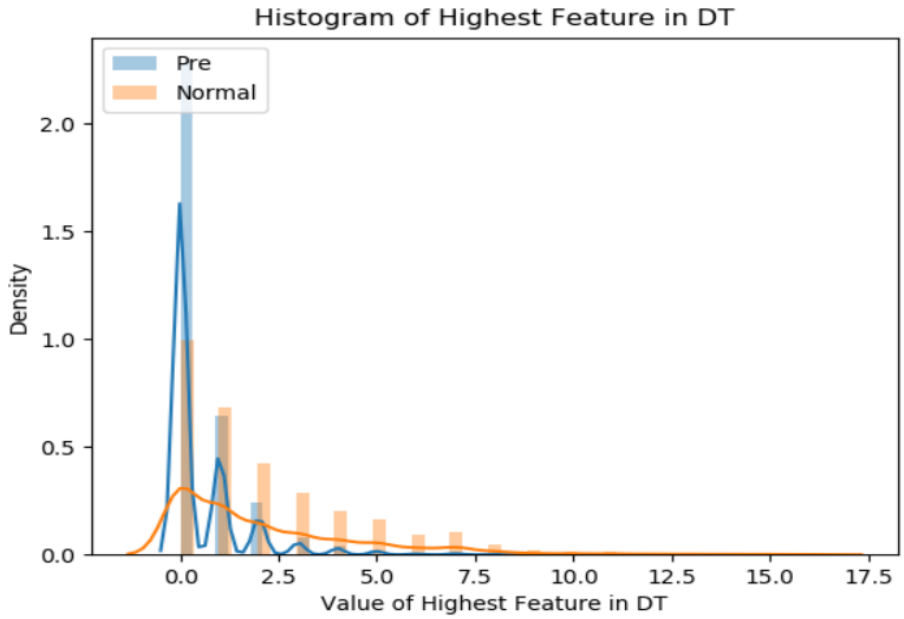

Figure 4-17: Tweet-level histogram of highest feature for pre-diagnosed group vs. control group.

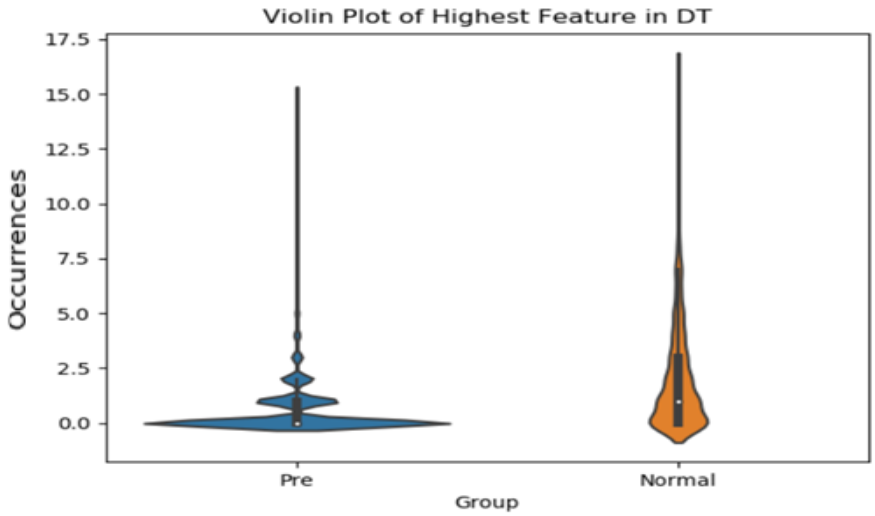

Figure 4-18: Tweet-level violin plot of highest feature for pre-diagnosed group vs. control group.

In figure 4-18, the violin plot shows the density of data at different values. The median value (represented by the white dot in the middle) for the control group is greater than 0 and is 0 for the pre-diagnosed group. The black bar is the interquartile range. The broader sections of the plot represent the occurrences of the highest feature in the pre-diagnosed group and the post-diagnosed group.

4.4.3 Post-diagnosed group vs. Control group

The decision tree in figure 4-19 shows the tweet-level class separation between the post-diagnosed group and the control group. The dependent variable of the decision tree (the root) has 15379 observations and two classes, true or false. The highest feature is feature 24778 in the dataset. This corresponds to the part of the speech '*superlative adjective*'. Similarly, the feature 24767 is '*disgust*', feature 24781 is '*singular noun'*. The F1 Score of the decision tree is 69%.

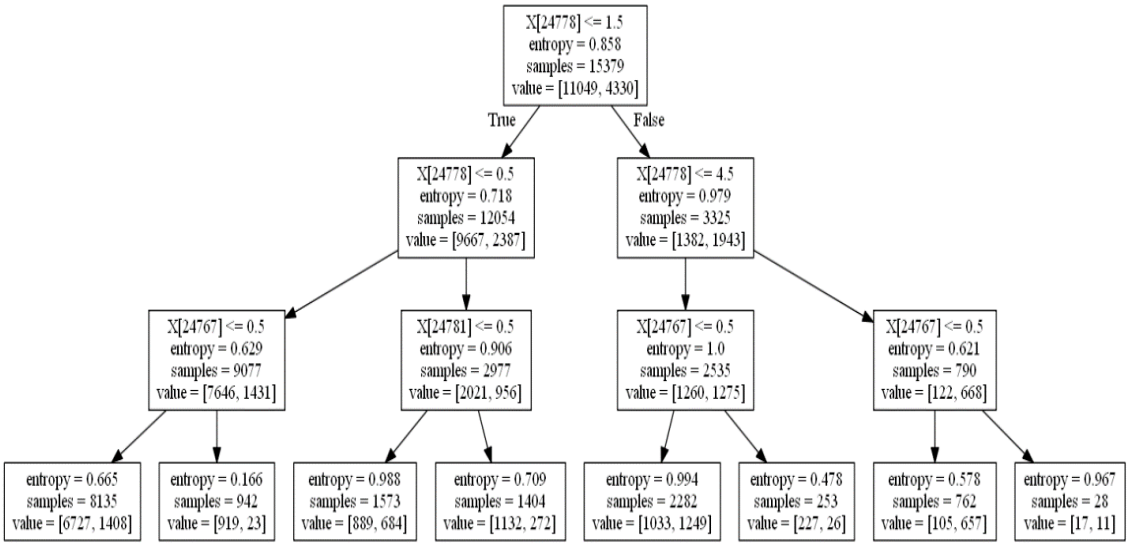

Figure 4-19: Tweet-level decision tree for post-diagnosed group vs. control group.

In figure 4-20, the label normal in the histogram represents the control group. The histogram for the post-diagnosed vs. control group on a tweet level is skewed right and therefore is asymmetrical. The graph shows a higher density for post-diagnosed if the feature value is 0. The control group has a higher density for values greater than 0 and less than 7.5. The density for the post-diagnosed group for feature values between 5 and 7.5 is 0.

Figure 4-21 shows the values of the highest feature. The median value (represented by the white dot in the middle) for the control group is greater than 0 and is 0 for the post-diagnosed group. The black bar is the interquartile range. The broader sections of the plot represent the occurrences of the highest feature in the pre-diagnosed group and the post-diagnosed group.

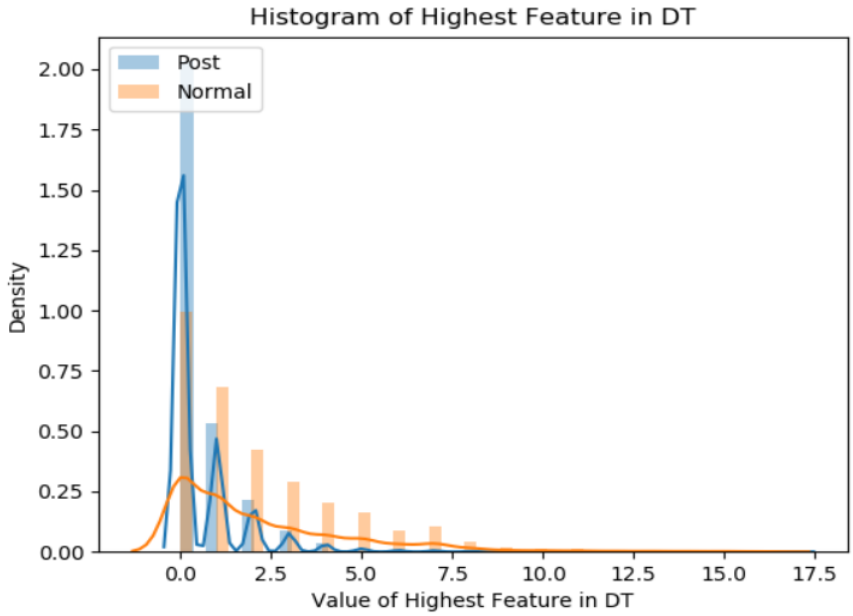

Figure 4-20: Tweet-level histogram of highest feature for post-diagnosed group vs. control group.

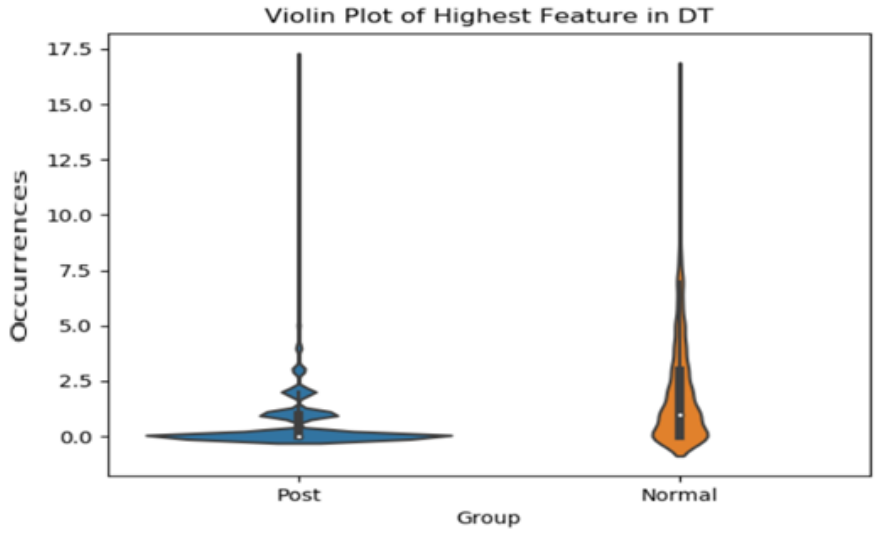

Figure 4-21: Tweet-level violin plot of highest feature for post-diagnosed group vs. control group.

4.4.4 Pre-diagnosed group vs Post-diagnosed group vs Control group

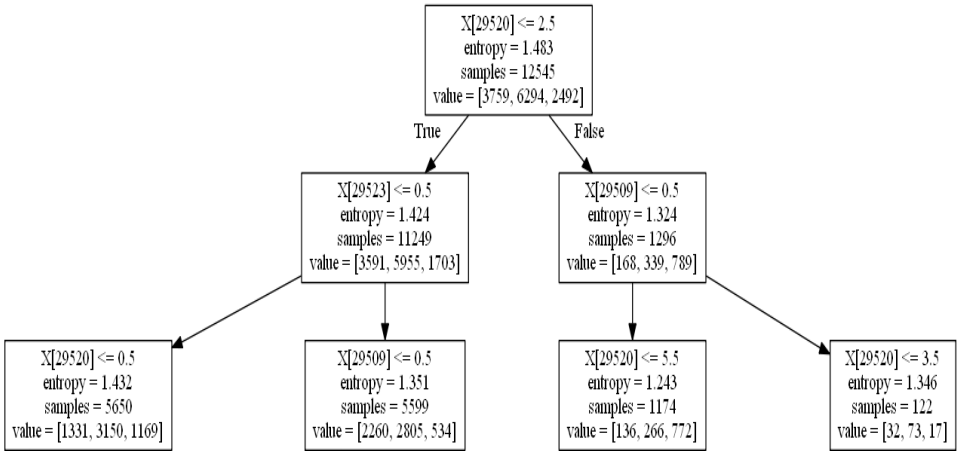

Figure 4-22: Decision tree for pre-diagnosed group vs post-diagnosed group vs. control group.

The train test size for this analysis is 60/40, and the argument average for the f1-score is. This was done to obtain the best possible accuracy.

The decision tree in figure 4-22 shows the tweet-level class separation between the pre-diagnosed group, the post-diagnosed group, and the control group. The dependent variable of the decision tree (the root) has 12545 observations and two classes, true or false. The highest feature is feature 29520 in the dataset. This corresponds to the part of the speech '*superlative adjective'*. Similarly, the feature 29523 is '*singular noun*', feature 29509 is '*disgust*. The F1 Score of the decision tree is 54%.

In figure 4-23, the label normal in the histogram represents the control group. The histogram for the post-diagnosed vs. control group on a tweet level is skewed right and therefore is asymmetrical. The graph shows a higher density for post-diagnosed and pre-diagnosed if the feature value is 0. If the feature value is between 0 and 2.5, control and pre-diagnosed have fared better. The control group has a higher density for values higher than 2.5 and less than 7.5.

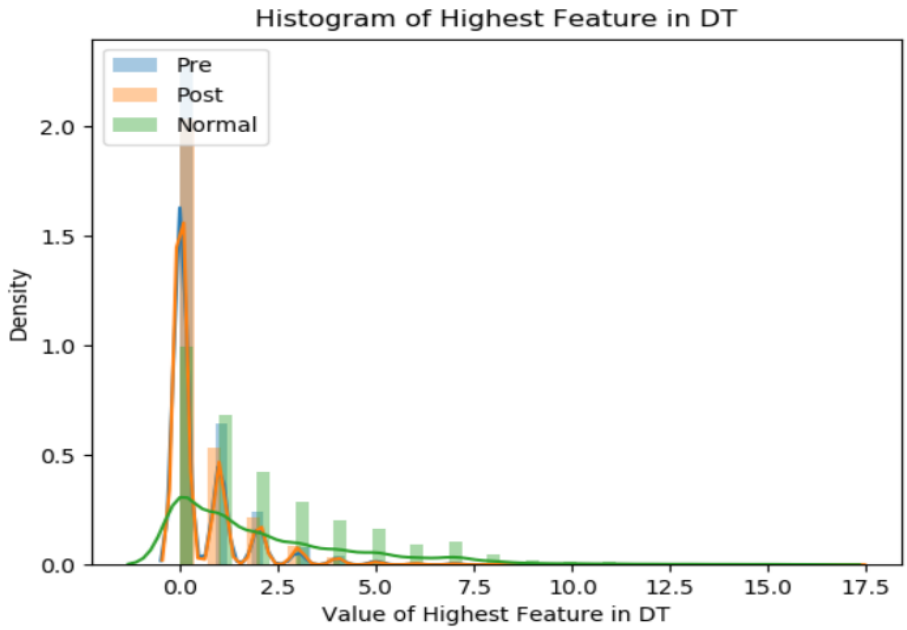

Figure 4-23: Tweet-level histogram of highest feature for pre-diagnosed group vs. post-diagnosed group vs. control group.

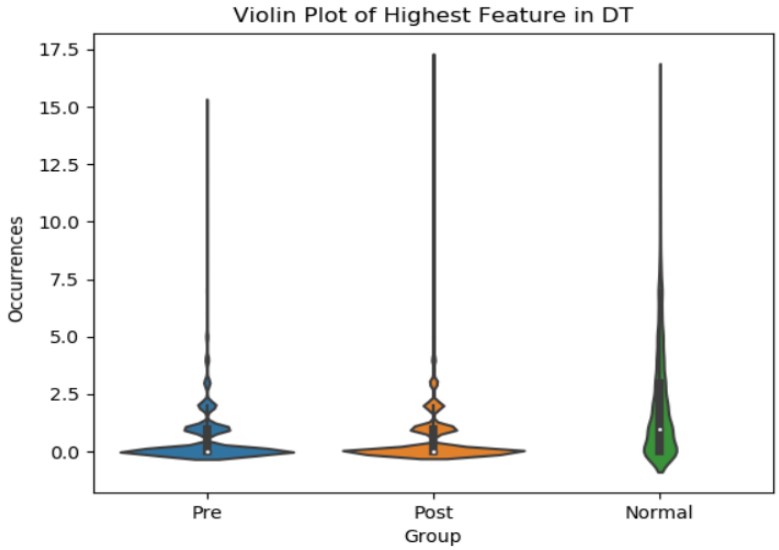

Figure 4-24: Tweet-level violin plot of highest feature for pre-diagnosed group vs. post-diagnosed group vs. control group.

In figure 4-24, the violin plot shows the values of the highest feature. The median value (represented by the white dot in the middle) for the control group is greater than 0

and is 0 for the pre-diagnosed and post-diagnosed group. The black bar is the interquartile range. The broader sections of the plot represent the occurrences of the highest feature in the pre-diagnosed group and the post-diagnosed group.

4.5 F1-Score

The table 4-3 lists the f1-score for the user-level analysis and the table 4-4 lists the f1-score for the user-level analysis:

Table 4-3: F1-score for user-level analysis.

Case	F1-Score
Pre-diagnosed vs. Post-diagnosed	0.80
Pre-diagnosed vs. Control	0.79
Post-diagnosed vs. Control	0.93
Pre-diagnosed vs. Post-diagnosed vs. Control	0.70

Table 4-4: F1-score for tweet-level analysis.

Case	F1-Score
Pre-diagnosed vs. Post-diagnosed	0.76
Pre-diagnosed vs. Control	0.72
Post-diagnosed vs. Control	0.69
Pre-diagnosed vs. Post-diagnosed vs. Control	0.54

4.6 Conclusion

The chapter explores the development of a clinical decision support system for behavioral disorders. The use of the decision tree is successfully able to distinguish between users in each of the groups. The decision trees in the four cases show the highest feature and its associated density. Similarly, the final accuracy of the classifier is dependent on how often it can clearly distinguish between an ADHD patient and a non-ADHD patient.

CHAPTER 5

CONCLUSION AND FUTURE WORK

The goal of this dissertation was the development of a clinical decision support system to assist in the diagnosis of individuals with Attention Deficit Hyperactivity Disorder. The clinical decision support system is based on three behavioral measures for the disorder. These measures are based on sentiment and semantics: variations in phrase structure rules, topic detection, sentiment, and emotion. Three groups of users, namely, the pre-diagnosed group, the post-diagnosed group, and the control group, form the classes to show differences in users before diagnosis and after.

The overarching objective was to support clinical decision making using a computational framework. To attain this, regression, unsupervised, and supervised approaches to the model locality were employed. That allowed us to uncover previously unknown and potentially useful information and finally support diagnosis automatedly by accessing social media. The supervised learning performed better than unsupervised learning for topic detection.

The clinical decision support system's applicability is generic and may apply to other behavioral disorders. With the help of the support system, the diagnosis and treatment of the disorder outside the United States may be made possible.

5.1 Conclusions

5.1.1 Correlations in language and emotion by the geographical prevalence

The correlations in language and emotion by geographic prevalence are established by using regression and cross-validation. The incidence of emotion, socio-economic status and emotion, and socio-economic status are calculated using Pearson product momentum correlation. Emotion and Socio-economic status outperform all other features. T-SNE and DBSCAN are used to cluster the three groups by geographical prevalence. The categories used for clustering are prevalence, anger, anxious, disengagement, engagement, negative emotions, positive emotions, negative relationships, and positive relationships.

5.1.2 Behavioral measures of Attention Deficit Hyperactivity Disorder

The symptoms of Attention Deficit Hyperactivity Disorder, inactivity; hyperactivity; and impulsivity have been quantified using three behavioral measures, as mentioned above: variations in phrase structure rules, topic detection, sentiment, and emotion. To establish variations in phrase structure rules, the collected tweets are broken down into their respective parts of speech. The elements of speech tags are based on the tag categories from the PennTree Bank. Rules are constructed for the three groups of users to show variations in speech.

The second behavioral measure, topic detection, have been found using Term frequency-inverse document frequency (TF-IDF) matrices and non-negative matrix factorization. The top ten topics for each of the three groups have been listed in chapter 3. The last behavioral measure contains the sentiments and emotions most commonly expressed by users of the three groups. The categories for the sentiments and emotions are based on NRC Emoticon.

The three behavioral measures are consistent with the symptoms and characteristics of the disorder. The sentiment most expressed by the pre-diagnosed group is negative as compared to the post-diagnosed group and control group, where the sentiment most expressed is positive. The sentiment for the pre-diagnosed group highlights a user's tendency to be easily excited (hyper) and argue their part. The nature of behavioral disorders also leaves individuals with feelings of resentment and social inadequacy. The parts of speech most commonly used by the three groups of users reflect the use of singular nouns for the pre-diagnosed and post-diagnosed group, and a plural noun or singular noun for the control group. The usage of singular nouns by users in the diagnosed group is symptomatic with their ability to only focus on a person, event, or thing at a time. The results for the behavioral measures can be translated into a questionnaire as a first step screening process in diagnosing individuals with Attention Deficit Hyperactivity Disorder.

5.1.3 Clinical decision support system for behavioral disorders

The classification of a user into one of the three classes provides the clinical decision support, making it the need of the hour. It provides a clinician with the information to be able to support their end decision regarding whether a patient has ADHD or not. The use of social media in this approach may help clinicians reach areas where ADHD is not considered to be a mental disorder.

The three behavioral measures are used as input for a decision tree classifier. The classification is on a tweet level and a user level. Four cases are of classification are considered: pre-diagnosed vs. post-diagnosed, pre-diagnosed vs. control group, post-diagnosed vs. control group, and pre-diagnosed vs. post-diagnosed vs. control group. The

accuracy of the classifier is better for the user-level analysis. The accuracy for the first case is 80%, the second is 79%, the third is 93%, and the last case is 70%.

5.2 Future Work

Future applications of the decision support classifier include its applicability to other behavioral disorders. It has the potential to answer questions related to the disorder, such as time to remission, the peak of the disorder, the type of the disorder. The type of disorder is the weight of inattention, hyperactivity, and impulsivity.

BIBLIOGRAPHY

Attention-deficit/hyperactivity disorder. (2019). [Online] Available at: https://www.mayoclinic.org/diseases-conditions/adhd/symptoms-causes/syc-20350889

Balusu M., Meghani T., Eisenstein J. (2018). Stylistic Variation in Social Media Part-of-Speech Tagging. *Proceedings of the Second Workshop on Stylistic Variation*. New Orleans, LA.

Brown G.W., Bifluco A., Veiel H.O.F., Andrews B. (1990). Self Esteem and Depression. *Aetiological Issues, Social Psychiatry and Psychiatry Epidemiology,* Volume 25, pp. 235-243.

Brownstein J.S., Freifeld C.C., Madoff L.C. (2009). Digital Disease Detection - Harnessing the Web for Public Health Surveillance. *New England Journal of Medicine,* pp. 2153-2157.

Carchiolo V., Longheu A., Malgheri M., Mangioni G. (2015). Multisource agent-based Healthcare Data Gathering. *Proceedings of the Federated Conference on Computer Science and Information Systems*.

CDC - Data and Publications. (2018). [Online] Available at: https://www.cdc.gov/mentalhealth/data_publications/index.htm

CDC - Mental Health. (2019). [Online] Available at: https://www.cdc.gov/childrensmentalhealth/symptoms.html

Cheong F., Cheong C. (2011). Social Media Data Mining: A Social Network Analysis of Tweets During the 2010-2011 Australian Floods. *Pacific Asia Conference on Information Systems.*

Chew C., Eysenbach G. (2010). Pandemics in the age of Twitter: Content Analysis of Tweets During the 2009 H1N1 Outbreak. *PloS ONE,* Volume 5(11).

Cloninger C.R., Svrakic D.M., Pryzbeck T.R. (1993). *A* psychological model of temperament and disorder. *Archives of General Psychiatry,* Volume 50(12), pp. 975-990.

Cloninger C.R., Svrakic D.M., Pryzbeck T.R. (2006). Can Personality Assessment Predict Future Depression? A Twelve-Month Follow-up of 631 Subjects. *Journal of Affective Disorders,* Volume 92, pp. 35-44.

Coppersmith G., Dredze M., Harman C., Hollingshead K., Mitchell M. (2015). *CL* Psych 2015 Shared Task: Depression and PTSD on Twitter. *Proceedings of the Second Workshop on Computational Linguistics and Clinical Psychology.* Denver, CO.

Coppersmith G., Harman C., Dredze M. (2014). Measuring Post Traumatic Stress Disorder in Twitter. *Association for the Advancement of Artificial Intelligence*.

Data and Statistics about ADHD. (2019). [Online] Available at: https://www.cdc.gov/ncbddd/adhd/data.html

DBSCAN. (2019). [Online] Available at: https://en.wikipedia.org/wiki/DBSCAN

De Choudhury M., Counts S., Horvitz E. (2013a). Predicitng Postpartum Changes in Emotion and Behavior via Social Media. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.*

De Choudhury M., Gamon M., Counts S., Horvitz E. (2013b). Predicting Depression via Social Media. *Association for the Advancement of Artificial Intelligence*.

Durme V. (2012). Streaming Analysis of Discourse Participants. *Proceedings of Empirical Methods in Natural Language Processing.*

Eichstaedt J.C., Schwartz H.A., Kern M.L., Park G., Labarthe D.R., Merchant R.M., Jha S., Agrawal M., Dziurzynski L.A., Sap M., Weeg C., Larson E.E., Ungar L.H., Seligman M.E. (2015). Psychological Language on Twitter Predicts County-Level Heart Disease Mortality. *Association for Psychological Science,* Volume 26(2), pp. 159-169.

Emotional and Behavioral Disorder. (2019). [Online] Available at: https://www.gadoe.org/Curriculum-Instruction-and-Assessment/Special-Education-Services/Pages/Emotional-and-Behavioral-Disorder.aspx

Ester M., Kreigel H.P., Sander J., Xu. (2003). [Online] Available at: http://www.cs.fsu.edu/~ackerman/CIS5930/notes/DBSCAN.pdf

Ginsberg J., Mohebbi M.H., Patel R.S., Brammer L., Smolinski M.S., Brilliant L. (2009). Detecting Influenza Epidemics Using Search Engine Query Data. *Nature,* Issue 457, pp. 1012-1014.

Han J., Kamber M., Pei J. (2000). Data Mining: Concepts and Techniques.

Hulkower R. (2016). Treating Attention-Deficit/Hyperactivity Disorders in Children Under Age Six Years: A Research Anthology. *Center for Disease Control*.

Ji X., Chun S.A., Wei Z., Geller J. (2015). Twitter Sentiment Classification for Measuring Health Concerns. *Social Network Analysis and Mining*.

Kilyeni A. (2014). Likes, Tweets and Other "Friends": Social Media Buzzwords From A Terminology Perspective. *Second Global Conference on Linguistics and Foreign Language Teaching*. Dubai, UAE.

Kotikalapudi R., Chellappaki S., Montgomery F., Wunsch D., Lutzen K. (2012). Associating Internet Usage with Depressive Behavior Among College Students. *Digital Object Identifier*.

Kullback-Leibler Divergence. (2019). [Online] Available at: https://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler_divergence

Lu Y., Zhang P., Liu J., Li J., Deng S. (2013). Health-Related Hot Topic Detection in Online Communities Using Text Clustering. *PLOS ONE*.

Meftah S., Semmar N., Sadat F., Raaijmakers S. (2018). Using Neural Transfer Learning for Morpho-syntactic Tagging of South-Slavic Languages Tweets. *Proceedings of the fifth Workshop on NLP for Similar Languages, Varieties, and Dialects*. Santa Fe, New Mexico.

Moreno M.A., Jelenchick L.A., Egan K.G., Coz E., Young H., Gannon K.E., Becker T. (2011). Feeling bad on Facebook: Depression Disclosures by College Students on a Social Networking Site. *Depression and Anxiety,* Volume 28(6), pp. 447-455.

Ordinary Least Squares. (2019). [Online] Available at: https://en.wikipedia.org/wiki/Ordinary_least_squares

Paul M.J., Dredze M. (2011). You Are What You Tweet: Analyzing Twitter for Public Health. *Association for the Advancement of Artificial Intelligence*.

Pearson Correlation Coefficient. (2019). [Online] Available at: https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

Qadir A., Mendes P., Gruhl D., Lewis N. (2015). Semantic Lexicon Induction from Twitter with Pattern Relatedness and Flexible Term Length. *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Robinson M.S., Alloy L.B. (2003). Negative Cognitive Styles and Stress-Reactive Rumination Interact to Predict Depression: A Prospective Study. *Cognitive Therapy and Research,* Volume 27(3), pp. 275-291.

Rude S., Gortner E.V., Pennebaker J. (2004). Language Use of Depressed and Depression-Vulnerable College Students. *Cognition and Emotion*, pp. 1121-1133.

Rude S., Valdez C.R., Odom S., Ebrahimi A. (2003). Negative Cognitive Biases Predict Subsequent Depression. *Cognitive Therapy and Research*, Volume 27(4), pp. 415-429.

Schneider J. (1997). Cross Validation. [Online] Available at: https://www.cs.cmu.edu/~schneide/tut5/node42.html

Seifter A., Schwarzwalder A., Geis K., Aucott J. (2010). The Utility of "Google Trends" for Epidemiological Research: Lyme disease as an example. *Geospatial Health*, Volume 4, pp. 135-137.

Surian D., Nguyen D.Q., Kennedy G., Johnson M., Coiera E., Dunn A.G. (2016). Characterizing Twitter Discussions about HPV Vaccines Using Topic Modeling and Community Detection. *Journal of Medical Internet Research*, Vol 18(No 8).

t-Distributed Stochastic Neighbor Embedding. (2019). [Online] Available at: https://en.wikipedia.org/wiki/T-distributed_stochastic_neighbor_embedding

TF-IDF. (2019). [Online] Available at: https://en.wikipedia.org/wiki/Tf%E2%80%93idf

Volkova S., Bachrach Y., Armstrong M., Sharma V. (2015). Inferring Latent User Properties from Texts Published in Social Media. *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Volkova S., Coppersmith G., Durme V. (2014). Inferring User Political Preferences from Streaming Communications. *Proceedings of ACL*.

Wang X., Wei F., Liu X., Zhou M., Zhang M. (2011). Topic Sentiment Analysis in Twitter: A Graph-based Hashtag Sentiment Classification Approach. *CIKM '11*. Scotland, UK.

Witten I., Frank E., Hall M., Pal C. (2016). In: Data Mining: Practical Machine Learning Tools and Techniques, pp. 177-178.

Witten I., Frank E., Hall M., Pal C. (2016). In: Data Mining: Practical Machine Learning Tools and Techniques, pp. 151-152.

Witten I., Frank E., Hall M., Pal C. (2016). In: Data Mining: Practical Machine Learning Tools and Techniques, pp. 119-121.

Zamal F.A., Liu W., Ruths D. (2012). Homophily and Latent Attribute Inference: Inferring Latent Attributes of Twitter Users from Neighbors. *Proceedings of te International AAAI Conference on Web and Social Media* .