Catch Me if You Can: Detecting Bipolar Disorder from Social Media

Daniel Anadria, s3091678

Faculty of Behavioural and Social Sciences, University of Groningen

Supervisors: Dr. L.F. Bringmann and Dr. G. Bouma

PSB3E-BTHO: Bachelor Honours Thesis

30-06-2021

This study was preregistered online at https://osf.io/qstej

Author note:

I am grateful to Laura Bringmann and Anna Langener for guiding me and providing me with their great insights at numerous times throughout this project. Also to Gosse Bouma for his time and support of the project. Additionally, I would like to thank Carmen Zürcher who helped and listened when the code did not. Furthermore, my thanks go to the IR lab team at Georgetown University for creating and sharing the Self-reported Mental Health Diagnosis dataset.

Finally, I thank the Center for Information Technology of the University of Groningen for their support and for providing access to the Peregrine high performance computing cluster.

Abstract

Recent studies from the intersection of computational linguistics and psychology claim high classification accuracy in the prediction of mental health disorders from user-generated text on social media. While such studies hold great promise for early detection and intervention in mental health conditions, they generally lack robustness checks that would determine whether the algorithms truly detect specific disorders, or capture mental health problems in general. Using the Self-reported Mental Health Diagnosis (SMHD) dataset created by Cohan et al. (2018), this study examines whether natural language processing and machine learning techniques applied to a large collection of Reddit user posts can be used to classify bipolar disorder and distinguish it from other mental health conditions, namely depression, a common differential diagnosis for bipolar disorder, and autism, a distinct mental health condition. A number of user behavior features are extracted from the dataset to be used as predictors, as well as psycholinguistic features based on LIWC 2015, and a TF-IDF-weighted bag of words. Three predictive algorithms are used for binary prediction of bipolar disorder: logistic regression, support vector machine and random forest. There are three specifications for the algorithms: (1) training and testing on users with bipolar disorder and their healthy controls, (2) training in the first specification but testing on the sample containing users with bipolar, autism, depression and healthy controls and (3) training and testing on users with bipolar, autism, depression and healthy controls. The best performance is observed in the first specification, while performance in the second and third specification is markedly lower. Users with depression and autism are captured at similar rates as users with bipolar, indicating that the algorithms may be capturing a broader spectrum of mental health disorders than mood disorders (such as bipolar and depression) or bipolar specifically. The paper highlights the importance of including robustness checks in the stream of research. More work needs to be done to understand how to capture specific mental health conditions such as bipolar disorder.

Keywords: bipolar disorder, social media, machine learning, natural language processing

Catch Me if You Can: Detecting Bipolar Disorder from Social Media

Globally, around 792 million people or 10.7% of the world population suffer from mental health disorders (Ritchie & Roser, 2018). In the European Union, this number is one in six (OECD & European Union, 2018). A rise in the number and severity of mental illnesses is expected in light of the COVID-19 pandemic (UN, 2020). Most mental health disorders start in adolescence (Patel et al., 2007), half of which manifest by age 14 and 75% are established by age 24 (Kessler et al., 2005). Early onset is linked to longer illness duration and higher comorbidity with other disorders (Caspi et al., 2020). Early diagnosis and intervention both show great promise for improvement of long-term outcomes of mental illness (Jiang et al., 2020), yet disorders often remain undetected and untreated until later in life (Angst et al., 2011; Williams et al., 2017).

Mental health disorders are often difficult to diagnose in clinical settings and many mental health disorders go undiagnosed, with, for example, as many as 50% of depression cases remaining undetected (Paykel et al., 1997). Additionally, comorbidity of different psychiatric disorders combined with their internal heterogeneity present a challenge for diagnosis and treatment of many conditions (Phillips & Kupfer, 2013). Close to 70% of bipolar patients are initially misdiagnosed (Lish et al., 1994) and on average, it takes between 5 to 10 years to reach the correct diagnosis (Baldessarini et al., 2007; Phillips & Kupfer, 2013).

Social media opens up opportunities for early diagnosis and prevention of mental health disorders. Recent studies have produced promising results in detecting mental health disorders from texts users shared on social media (De Choudhury et al., 2013; Coppersmith et al., 2014; Shen and Rudzicz, 2017; Sekulić et al., 2018). Such studies rely on machine learning (ML) and natural language processing (NLP) algorithms to detect signals in large amounts of textual data. This approach primarily targets young adults who are the key demographic represented on social media platforms (Amir et al., 2019). The aim of these studies is to find unique linguistic markers of specific mental health disorders from text users share on social media platforms (Jiang et al., 2020). Once the affected users are identified, interventions can be planned. However, the diagnostic value of ML and NLP algorithms at predicting mental health from text remains to be scrutinized.

Studies predicting mental health disorders from user-generated text on social media have generally achieved high accuracy in binary classification tasks, in other words in predicting whether or not a user suffers from a specific condition. For example, in their study predicting bipolar disorder on Reddit, Sekulić et al. (2018) achieved high accuracy and F1 score of 0.86. Similar results have been found for depression (De Choudhury et al., 2013), PTSD (Coppersmith et al., 2014), schizophrenia (Mitchell et al., 2015) and various other disorders (Cohan et al., 2018). However, most studies lack robustness checks to determine whether the algorithms truly detect a specific disorder or just mental health disorders in general. In their sample containing nine mental health disorders, Cohan

et al. (2018) were able to achieve good precision, recall and F1 values on binary classification tasks. However, the algorithms struggled to recognize specific disorders in multi-label multi-class setting where they were tasked with prediction of a plethora of disorders given a user.

The present study builds on the previous work in the cross-section of computational linguistics and mental health prediction using Cohan et al. (2018) Self-reported Mental Health Diagnosis (SMHD) dataset which is further described in the Data and Methods section of this paper. The aim of the present work is to establish whether ML and NLP algorithms can truly detect bipolar disorder from user-generated text on Reddit, and if they are able to differentiate bipolar from other mental health disorders, namely depression and autism. Like bipolar disorder, depression is a mood disorder and a common differential diagnosis for bipolar in the DSM-5 (American Psychiatric Association, 2013) and as such may provide a challenge for the algorithms similar to what a clinician might face. Autism is a neuro-developmental disorder that is not a differential diagnosis for bipolar disorder. Therefore, it is quite distinct and a clinician would likely not confuse the two, and the question arises of whether an algorithm would.

The research question this paper attempts to answer is whether it is possible to differentiate bipolar disorder from depression, autism and healthy controls through identification of unique linguistic markers from large collections of Reddit post and comment data via ML and NLP techniques? Furthermore, this study investigates whether algorithms are actually predicting bipolar from user-generated text or whether user behavior produces the predictive power.

Related work

This section summarizes most relevant articles related to the present work. Since there is a plethora of papers using ML and NLP techniques to predict and describe mental health disorders from user language, this section presents a summary of most seminal and representative works.

De Choudhury et al. (2013) examined depressed users on Twitter. The authors conducted a study using Amazon's Mechanical Turk interface where they issued Center for Epidemiologic Studies Depression (CES-D) scale questionnaires to participants followed by questions on participant depression history and demographics. The final question asked the participants to share their public Twitter profiles which were then used for further analysis. The authors developed a support vector machine classifier capable of detecting depressed users with accuracy of more than 70% and precision above 80%. Furthermore, they identified Twitter user behaviors related to depression. Namely, depressed users were more likely to be active late at night compared to their non-depressed counterparts. Furthermore, depressed users made overall less posts and replies. The Linguistic Inquiry and Word Count (LIWC) tool, a text analysis program discussed in Data and Methods, showed more negative affect and lower activation relative to non-depressed controls, as well as higher incidence of

first-person pronouns such as “I”, “me”, and “my”, and lower use of third person pronouns such as “he”, “she”, “his”, and “her”. The findings are in line with depression literature stating that depressed individuals are more likely to be active at night, show a decrease in social connection, higher negative affect, and increased self-focus (De Choudhury et al., 2013).

Coppersmith et al. (2014) examined the language use of Twitter users affected by post-traumatic stress disorder (PTSD). The users were identified through an automated search for strings related to PTSD diagnosis and later manually inspected. The affected users were matched with healthy controls. LIWC categories were used to inspect differences in language use between PTSD users and healthy controls. The use of second-person pronouns such as “you” and “yours” was found to be more rare in PTSD users. Unlike users with depression, users with PTSD showed an increase in third-person pronoun use and words expressing anxiety. The authors trained several loglinear regression classifiers using features obtained from a unigram language model, a character n-gram language model, and one composed from LIWC categories. Unigram models focus on words in isolation while n-gram models consider *n* words in sequence (Zheng & Casari, 2018). The unigram language model showed best performance while authors note that LIWC did not perform as well at the binary prediction task.

Sekulić et al. (2018) studied prediction of bipolar disorder from user-generated text on Reddit. The authors collected posts from subreddits related to bipolar disorder and applied an automated search for self-reported diagnosis strings. After processing, the yielded dataset contained 3,488 bipolar Reddit users. The healthy control users were sampled from subreddits – topic-based online communities on Reddit – often visited by bipolar users. The authors created a balanced dataset with 3,931 healthy controls. Three sets of features were computed per user: psycholinguistic features using LIWC and Empath (a tool similar to LIWC), lexical features using a TF-IDF-weighted bag of words, and Reddit user behavior features such as post-to-comment ratio, number of awarded posts and average controversiality. In binary prediction of bipolar disorder, three prediction algorithms were used: support vector machine (SVM), logistic regression, and random forest ensemble (all three of which are explained in the Data and Methods section of this paper). Out of these, random forest ensemble achieved the highest accuracy and F1 score of 0.869 and 0.863, respectively. Similar to depressed users, the use of first-person pronouns was more prevalent in the bipolar group, as well as authentic speech, words reflecting emotions, health and biological processes. Bipolar users used more words related to both positive emotions and sadness while healthy controls reflected more anger in their writing.

Cohan et al. (2018) created a large labeled SMHD dataset for studying the presence of nine different mental disorders on Reddit. These are attention deficit hyperactivity disorder, anxiety disorders, autism, bipolar disorder, depressive disorders, eating disorders, obsessive-compulsive

disorder, post-traumatic stress disorder and schizophrenia. LIWC was used to explore psycholinguistic differences between each disorder and healthy controls. Clout or the language indicating high social status was more prevalent among control users compared to depressed and anxious Redditors. Most mental disorders were characterized by increased authenticity scores indicating higher personal and first-person pronoun usage. The findings indicate that depressed and anxious users communicate in ways related to lower social status, and that higher language authenticity is a marker for mental disorders as the affected users tended to use more self-referential language. Depressed users scored higher on features relating to health, focus on the past, and biological processes, while controls scored higher on topics of money and leisure. The authors trained five classifiers both for binary prediction tasks, that is predicting whether a particular user has a specific condition, and multi-label multi-class setting which tries to predict which disorders if any from a plethora of conditions a user has been diagnosed with. Logistic regression had the highest precision in binary classification tasks for most disorders, and for most conditions convolutional neural network had the highest recall. The algorithms performed more poorly in a multi-label multi-class setting.

The above-mentioned studies indicate there may be meaningful differences in writing styles and user behaviors of people, allowing the prediction of specific mental health conditions. The present study puts this claim to a test.

Data and Methods

The SMHD dataset created by Cohan et al. (2018) contains users labeled by mental health conditions and all their Reddit posts between January 2006 and December 2017. All posts related to mental health as well as posts in mental health related subforums of Reddit, so called subreddits, were removed for both diagnosed users and healthy controls. The dataset is presplit into training, validation and testing subsets. Diagnosed users and their healthy controls are matched based on the subreddits they were active in. The authors developed high precision detection patterns to identify users with mental health conditions relying on diagnostic keywords, synonyms, common misspellings and several other inclusion criteria. A subsample of posts was manually inspected to assess the precision of classification which was estimated at 95.8%. Each diagnosed user was matched with on average nine control users based on a similar number of posts in the same subreddits. Although control users had twice the number of posts, they tended to be shorter when compared to diagnosed user posts (Cohan et al., 2018). While the original dataset contains users belonging to nine different disorders, the present analysis retains only those diagnosed with bipolar disorder, depression and autism, with other disorders represented only as comorbidities to the aforementioned. Table 1 shows the number of participants with autism, bipolar disorder and depression in each subset including their healthy

controls. Healthy controls for depression and autism are only used in the robustness checks of this paper.

Table 1

Number of diagnosed users and healthy controls in the SMHD

Disorder	Training	Validation	Testing	Total
Bipolar	683	679	735	2,097
with comorbidities other than depression and autism	335	308	312	955
with depression	187	187	189	563
with autism	11	8	11	30
Bipolar Total	1,216	1,182	1,247	3,645
Bipolar Healthy Controls	13,500	13,610	14,368	41,478
Depression	1,316	1,308	1,316	3,940
with comorbidities other than bipolar and autism	1,134	1,055	1,088	3,277
with autism	25	24	18	67
Depression Total	2,662	2,574	2,611	7,847
Depression Healthy Controls	19,921	19,652	20,219	59,792
Autism	348	340	361	1,049
with comorbidities other than bipolar and depression	95	108	127	330
Autism Total	479	480	517	1,476
Autism Healthy Controls	5,232	5,220	5,663	16,115
Healthy controls Total	38,653	38,482	40,250	117,385

As previously stated, the users who did not have at least one of the three conditions of interest, i.e. bipolar, depression or autism were removed from the dataset. When comorbidities were examined on this newly rendered dataset, 59.64% of users had none, followed by users with two conditions at 30.19%, three at 8.25%, and four or more disorders at 1.92%. Figure 1 shows correlations between the disorders in the dataset. The average number of matched healthy controls for diagnosed users can be seen in Table 2.

Figure 1

Correlation Matrix of Comorbidities

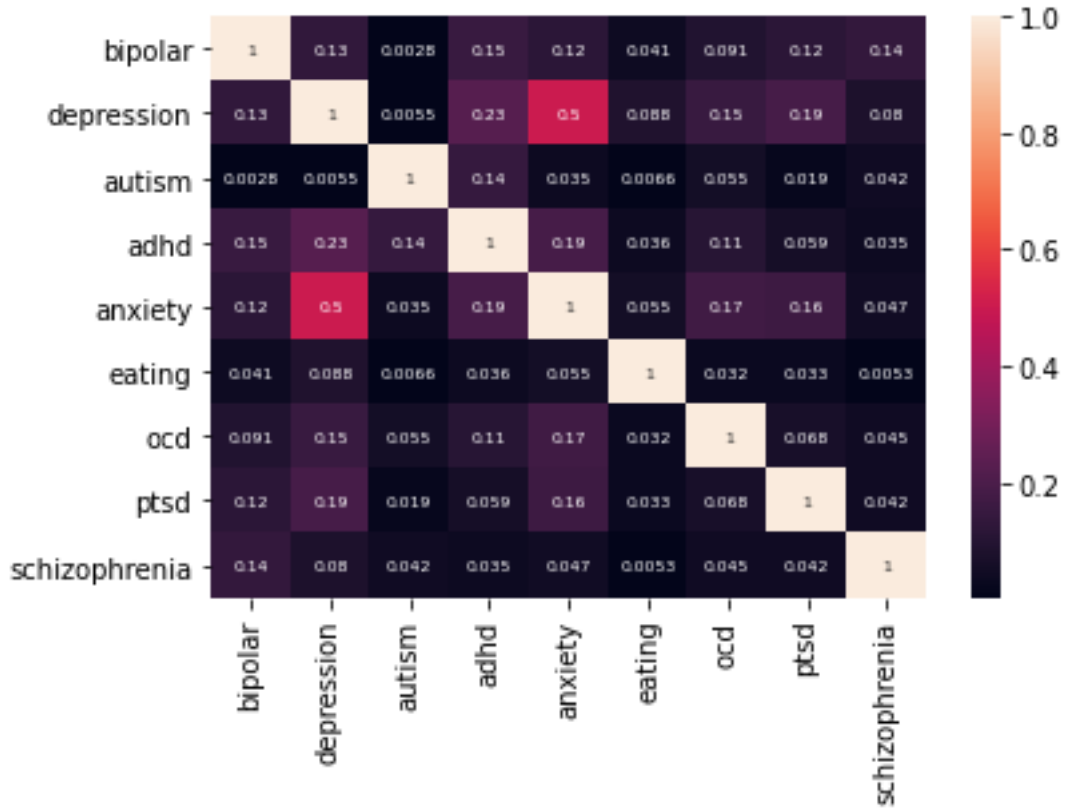

Table 2

The average number of matched healthy controls per diagnosed user

Disorder	Training	Validation	Testing
Bipolar	11.1	11.5	11.5
Depression	7.5	7.7	7.7
Autism	10.9	10.9	11

Feature extraction

A series of user behavior variables was extracted from the SMHD dataset. They are briefly described here and their means and standard deviations can be seen in Table 3.

Total activity represents the number of both posts and comments that users made with higher values indicating greater overall activity. The resulting total activity score was rounded to the nearest integer. The number of posts and comments made is generally larger for control users compared to the diagnosed users.

Comment ratio indicates what percentage of users' total activity is comments, as opposed to authored posts. Comment ratios are relatively uniform across subsets and conditions.

Posting interval is the average time elapsed between instances of user activity – posts or comments in days. Posting intervals appear higher for diagnosed users than their healthy controls.

Time active is the difference between a user's first and final post in the dataset computed in years. Time active shows that diagnosed users were active for slightly shorter amount of time than their healthy controls.

Post length is the average length of a post or comment in characters. Post lengths appear systematically longer for diagnosed users than their controls across all subsets.

Table 3

Means and standard deviations for the user behavior variables

Total activity	Training	Validation	Testing
Bipolar	163 (85)	155 (81)	155 (82)
Depression	165 (85)	161 (84)	160 (83)
Autism	162 (84)	175 (87)	168 (82)
Bipolar control	309 (159)	299 (153)	297 (154)
Depression control	318 (163)	313 (159)	307 (160)
Autism control	312 (158)	331 (169)	329 (158)
Comment ratio	Training	Validation	Testing
Bipolar	0.89 (0.11)	0.89 (0.10)	0.89 (0.11)
Depression	0.89 (0.11)	0.88 (0.11)	0.88 (0.12)
Autism	0.87 (0.15)	0.88 (0.14)	0.88 (0.13)
Bipolar control	0.88 (0.17)	0.89 (0.16)	0.88 (0.17)
Depression control	0.88 (0.16)	0.88 (0.11)	0.88 (0.16)
Autism control	0.88 (0.17)	0.88 (0.16)	0.88 (0.17)
Posting interval	Training	Validation	Testing
Bipolar	6.8 (6.5)	6.7 (6.4)	6.8 (6.6)
Depression	6.4 (5.8)	6.7 (6.3)	6.6 (6.2)
Autism	5.5 (5.1)	5.5 (6.0)	5.6 (5.7)
Bipolar control	4.6 (3.8)	4.5 (3.9)	4.7 (4.0)
Depression control	4.4 (3.7)	4.5 (3.8)	4.6 (3.9)
Autism control	4.3 (3.8)	4.1 (3.7)	4.2 (3.6)
Time active	Training	Validation	Testing
Bipolar	2.4 (1.8)	2.3 (1.8)	2.3 (1.8)
Depression	2.4 (1.7)	2.4 (1.8)	2.3 (1.8)
Autism	2.0 (1.6)	2.1 (1.7)	2.1 (1.7)
Bipolar control	3.1 (2.0)	3.1 (2.0)	3.1 (2.1)
Depression control	3.1 (2.0)	3.1 (2.0)	3.1 (2.0)
Autism control	3.0 (2.0)	2.9 (2.0)	3.0 (2.0)

Post length	Training	Validation	Testing
Bipolar	244 (207)	239 (165)	242 (186)
Depression	240 (143)	239 (165)	237 (157)
Autism	252 (193)	256 (160)	253 (164)
Bipolar control	139 (91)	141 (93)	140 (101)
Depression control	140 (100)	140 (86)	138 (93)
Autism control	142 (97)	146 (96)	144 (90)

Psycholinguistic and lexical features are extracted from text. This is done in two ways, (1) through the use of Linguistic Inquiry and Word Count (LIWC) 2015 lexicon and (2) through the use of a TF-IDF-weighted bag of words. Both approaches have been previously used in similar studies as discussed in the previous section (e.g. Sekulić et al., 2018) and are briefly described below.

LIWC 2015 is a text analysis program which takes in target words from a text file and scores each word and the overall text upon a plethora of psychological and linguistic scales. LIWC creates more than 90 output variables ranging from linguistic categories such as percentage of pronouns in the text, articles, auxiliary verbs and punctuation to psychological variables such as analytical thinking, emotional tone and affect. An individual word can belong to several categories. For example, the presence of the word “cried” will increase loadings of categories sadness, negative emotion, overall affect, verbs and past focus (Pennebaker et al., 2015a). In addition, LIWC has specialty dictionaries for analysis of texts written in various languages (Pennebaker et al., 2015b). The means and standard deviations for LIWC variables across mental health conditions can be found in the Appendix A.

Bag of words refers to a natural language processing approach where words from a text are represented in a vector with their corresponding frequency counts. The bag of words approach used in this study contains no sequence information, just single words with their frequencies (i.e. a unigram language model). A weakness of the bag of words approach is that when all the words are counted equally, some common words become overemphasized. These are called stop words. Stop words are some of the most common words in English such as articles, prepositions and pronouns that generally do not add a lot of value to the analysis as they tend to highly appear in all kinds of texts. A simple frequency count is therefore not enough for truly relevant words to stand out (Zheng & Casari, 2018). This is where TF-IDF weighting enters the stage.

Term Frequency-Inverse Document Frequency (TF-IDF) is a numeric measure expressing importance of a word in a given document. A high TF-IDF score indicates that a word is present in a few documents of the total corpus but tends to appear many times in the documents where it is present. Therefore, as a measure, TF-IDF is useful to determine keywords of a given document (Rajaraman & Ullman, 2011). The formula used to compute TF-IDF of a given word is a multiplication of term frequency (TF) and inverse document frequency (IDF), calculated as:

$$w_d = f_{w,d} * \log\left(\frac{D}{f_{w,D}}\right)$$

, where the TF-IDF score of a word *w* in a document *d*, which is part of document corpus *D* equals the term frequency *f* of word *w* in document *d*, multiplied by IDF or the log of the total number of documents in the corpus *D* divided by the number of documents the word *w* appears in $f_{w,D}$ (Ramos, 2003). TF-IDF prefers words that are frequent in the document *d* but are relatively rare in the entire corpus *D*. TF-IDF is often used as a way of weighting words in bag-of-words approaches where words are sampled without attention to their order while preserving the word frequency information (Jurafsky & Martin, 2021). This approach is applied in the present study.

In order to use **TF-IDF-weighted bag of words**, a number of changes need to be applied to the text to aid its further analysis. First, all the text is converted to lower case (e.g. *"*HELLO TOM*"* becomes *"*hello tom*"*). Next, stop words are removed. For example, in the sentence "A boy returned to his house", the words "boy", "returned" and "house" carry most of the semantic value of the sentence. Traditionally, words such as *"*a*"*, *"*to*"* and *"*his*"*, namely articles, prepositions and pronouns would be removed from the bag of words. These word categories have high frequency but carry little meaningful information which is why they are usually omitted. However, previous studies (e.g. De Choudhury et al, 2013; Coppersmith et al., 2014; Sekulić et al., 2018) have indicated some predictive value in the use of pronouns, especially when examining the writing styles of people with mental health disorders. Hence pronouns are retained in the present work regardless of their high frequency. Additionally, HTML syntax such as links shared by users are removed. This process keeps only alphanumeric values –letters and numbers. Finally, the words are stemmed. Stemming refers to the removal of suffixes that distinguish semantically similar words. For example, the words *"*buyer*"*, *"*buying*"* and *"*buy*"* are all stemmed to *"*buy*"* in order to be mapped as the same concept. In the present study, the words are stemmed using Porter stemmer, the most popular English language stemmer (Zheng & Casari, 2018). Finally, the TF-IDF-weighted bag of words vocabulary is constructed based on the training sample and words with document frequency less than 20 are removed from the analysis, similar to Cohan et al. (2018). The words that have undergone changes such as being switched to lower case, stopwords and HTML removal, and stemming are referred to as tokens and are used to compute TF-IDF frequencies based on the constructed vocabulary.

Prediction Techniques

In this subsection, the focus is on the machine learning techniques that will use both linguistic and user behavior features for prediction purposes. All classifier are trained for a binary classification task of bipolar vs. non-bipolar. Specifically, these are random forest, logistic regression and support vector machine algorithms. Hyperparameters for each model are chosen using the GridSearchCV function which allows for testing of different combinations of hyperparameter values at the same

time, also known as hyperparameter tuning, and models are tuned based on F1. The function uses a 5-fold cross-validation approach, which supports the optimization of the hyperparameters but also reduces the risk of overfitting. The train and validate sets are combined into a single set for the purposes of the analysis. Parameter grids for each algorithm are included in the Appendix B. The best performing model for each algorithm is chosen for the prediction task. For specifications including TF-IDF predictors, only logistic regression is used due to the high computational resource requirements.

Random forests (RF) are machine learning algorithms relying on ensembles of unique decision trees used for classification tasks. Each tree is randomized through selection of unique users and features with replacement, meaning the same collection can be sampled multiple times. This is called bootstrap sampling and it leads to each tree's uniqueness due to it being built on a slightly different dataset. A prediction is first made for each tree in the forest. In classification tasks, each tree provides a set of probabilities for each possible output. The probabilities are averaged across trees and the outcome with the highest overall probability is chosen. RF provide feature importance measures that can be used to compare feature performances in the forest. A greater number of trees within a random forest increases its robustness (Müller & Guido, 2016).

Logistic regression is a popular approach to classification tasks in which the response is a binary or discrete variable. It uses a logistic function to compute the probability of the outcome as a value between 0 and 1 The predictors can be categorical or continuous (James et al., 2013). Logistic regression is commonly taught in undergraduate psychology programs and therefore presents one of the most common approaches to binary classification tasks in the field of psychology. In the present work, logistic regression is regularized to avoid overfitting by adding a penalty term.

Support Vector Machine (SVM) is a machine learning algorithm used for classification tasks. It does so by separating the hyperplane containing the observations with a line of maximal distance from each observation. This separation of the space allows new observations to be classified depending on where they fall on the plane. In case of outliers, or observations that fall in the wrong class, SVM can be modified to have a soft margin which allows it to tolerate some data points on the wrong side of the margin. SVM's kernel function allows a mathematical transformation of non-linearly separable data in order to draw a separation line where it could otherwise not be drawn (Noble, 2006). Features are normalized using min-max scaling for SVM (Zheng & Casari, 2018).

The prediction task is always binary – bipolar vs. not bipolar. However, there are three specifications for the algorithms which refer to experimental setups, i.e. between which users the algorithms have to predict. An overview of the specifications is shown in Table 4.

Table 4

Overview of Specifications for Machine Learning Models

	1st Specification	2nd Specification	3rd Specification
Setup	Algorithms are trained and tested on users with bipolar disorder and healthy controls	Trained on 1st Specification, but tested on users with bipolar, autism, depression and healthy controls	Algorithms are trained and tested on users with bipolar, autism, depression and healthy controls
Rationale	Replicating the approach of previous studies	Testing the previous studies' approach on a real life scenario where other conditions exist	Training and testing on a real life scenario where other conditions exist

Findings and Discussion

The results for each specification are shown in Table 5 with S1, S2 and S3 referring to the first, second and third specifications respectively.

Table 5

Results of Machine Learning

Predictors	Algorithms		S1	S2	S3
User behavior only	RF	Precision	0.90	0.31	0.38
		Recall	0.35	0.35	0.04
		F1	0.51	0.33	0.07
	SVM	Precision	0.18	0.00	0.10
		Recall	0.21	0.00	0.12
		F1	0.19	0.00	0.11
	Logistic	Precision	0.75	0.31	0.38
		Recall	0.16	0.16	0.02
		F1	0.27	0.21	0.03
LIWC only	RF	Precision	0.80	0.30	1.00
		Recall	0.26	0.26	0.00
		F1	0.40	0.28	0.00
	SVM	Precision	0.78	0.00	0.18
		Recall	0.28	0.00	0.13
		F1	0.41	0.00	0.15
	Logistic	Precision	0.71	0.30	0.34
		Recall	0.38	0.38	0.04
		F1	0.50	0.34	0.07
User behavior + LIWC	RF	Precision	0.91	0.31	0.80
		Recall	0.47	0.47	0.00
		F1	0.62	0.37	0.01
	SVM	Precision	0.79	0.00	0.19
		Recall	0.33	0.00	0.16
		F1	0.47	0.00	0.17
	Logistic	Precision	0.72	0.30	0.37
		Recall	0.40	0.40	0.04
		F1	0.51	0.34	0.08
TF-IDF only	Logistic	Precision	0.82	0.33	0.50
		Recall	0.37	0.37	0.05
		F1	0.51	0.35	0.09
User behavior + TF-IDF	Logistic	Precision	0.81	0.32	0.54
		Recall	0.43	0.43	0.07
		F1	0.56	0.37	0.13

Note. Precision is equal to True Positives / (True Positives + False Positives), i.e. the share of true positives in the total number of predicted positives. Recall is equal to True Positives / (True Positives + False Negatives), i.e. the share of predicted true positives in the total true positives or how many of the bipolar users were correctly identified. F1 is equal to 2 * (Precision * Recall/(Precision + Recall), i.e. the harmonic mean of Precision and Recall which balances the two measures.

In Table 5, a general pattern is visible where performance is relatively high in the first specification, where the algorithms have to distinguish only between bipolar users and their healthy controls as in previous studies, but performance drops when other disorders are included in the second and third specifications. Regarding the first specification, when algorithms in this study are compared to Cohan et al. (2018), the present models show similar or better performance. However, as in Cohan et al. (2018), while precision is high, recall is fairly low with the highest score being 0.47 which means fewer than half of bipolar cases were identified. In the second specification, precision is reduced to less than half compared to the performance in the first specification indicating an increase in false positives. The third specification shows low overall performance, with the best model scoring an F1 value of 0.17 and precision of 0.19. The findings indicate that it is difficult to distinguish bipolar disorder in presence of other mental health disorders. Model performance for algorithms of the first and third specification in the train and validate set is similar to the performance in the test set which indicates that overfitting is not at play.

A closer examination of the predictions in the form of a confusion matrix in Table 6 shows that users with depression are classified as having bipolar at almost the same rate as bipolar users. The misclassification rate for autism is somewhat lower but still high. The confusion matrices for other specifications, algorithms and predictor sets yield a very similar picture. The rationale to including depression in the dataset was that it is a disorder which is a differential diagnosis for bipolar disorders, therefore providing a challenge for the algorithms. The findings are congruent with that. However, autism is clinically not a differential diagnosis for bipolar disorders – it is seen as a relatively distinct disorder. Hence, it could be expected from the algorithms to differentiate between autism and bipolar disorders at a higher rate. The evidence here shows only slight support to that proposition – the misclassification rates are somewhat lower than for depression, albeit still high.

Table 6

Confusion Matrix of Random Forest 2nd Specification for User Behavior + LIWC

True	Predicted		% classified as bipolar
	Bipolar	Not bipolar	
Bipolar	587	660	47.1
Depression	1062	1209	46.8
Autism	184	313	37.0
Healthy Control	56	14312	0.4

Furthermore, the findings in Tables 5 and 6 are robust to tuning based on recall rather than F1 score, exclusion of comorbidities, inclusion of depression and autism healthy controls and exclusion of any healthy controls (see Appendix C).

Next, the question of whether algorithms rely on user generated text or user behavior when predicting bipolar disorder is examined. The overview of the most important features for logistic regression and random forest can be found in Table 7. Overall, random forest appears to rely more on user behavior than logistic regression, as also supported by Table 5. When user behavior variables are not available (as in LIWC only), random forest loses about half of recall while logistic regression performs similarly for both LIWC and TF-IDF. The first and the third specification rely on similar features, but the third specification features have less predictive power. In summary, both textual and user behavior features are important predictors for distinguishing between mental health conditions and healthy controls, but may not be sufficient to predict bipolar disorder specifically.

Table 7

Feature importance overview for logistic regression and random forest

	User Behavior + LIWC				LIWC only			
	S1		**S3**		**S1**		**S3**	
Logistic Regression	anx	1.75	commentRatio	-0.85	anx	1.88	assent	-0.75
	health	1.29	anx	0.58	health	1.44	anx	0.67
	commentRatio	-1.25	body	-0.51	nonflu	-1.10	nonflu	-0.65
	nonflu	-0.83	discrep	-0.47	body	-0.89	body	-0.54
	we	-0.83	feel	-0.47	assent	-0.85	discrep	-0.50
	body	-0.72	bio	0.45	we	-0.79	feel	-0.48
	sexual	-0.57	assent	-0.44	anger	-0.66	bio	0.48
	discrep	-0.52	ingest	-0.43	ingest	-0.61	health	0.45
	ingest	-0.47	nonflu	-0.42	discrep	-0.59	ingest	-0.45
	anger	-0.45	sexual	-0.41	sexual	-0.52	netspeak	-0.43
Random Forest	totalActivity	0.15	totalActivity	0.04	health	0.06	health	0.04
	health	0.05	health	0.03	Quote	0.05	Quote	0.03
	postLength_mean	0.03	postLength_mean	0.02	anx	0.03	i	0.02
	anx	0.03	i	0.02	i	0.03	anx	0.02
	Quote	0.03	Quote	0.02	AllPunc	0.02	WC	0.02
	i	0.02	ppron	0.02	WC	0.02	AllPunc	0.02
	postLength_SD	0.02	anx	0.02	ppron	0.02	ppron	0.02
	ppron	0.02	postLength_SD	0.01	Comma	0.02	Comma	0.02
	WC	0.02	WC	0.01	conj	0.02	Dic	0.01

Overall, the findings indicate that while bipolar disorder can be detected with high precision given that it is the only disorder in the sample, the presence of other disorders leads to high misclassification rates. The high misclassification rates for autism further indicate that the algorithms may be capturing a broader spectrum of mental health disorders than mood disorders (such as bipolar and depression) or bipolar specifically. In other words, the algorithms seem to have difficulty distinguishing between bipolar and other mental health disorders based on the given sets of predictors. The algorithms were mostly able to distinguish between healthy controls and affected users with high

precision which indicates that mental illness might indeed have an effect on user writing style and behavior.

There are a number of limitations related to the SMHD dataset and this paper in general that are important to discuss. First, the number of user behavior variables that could be extracted from the dataset was limited. For example, it was not possible to determine users' time zones, even though late night activity might be a valuable predictor for bipolar disorder (Sekulić et al., 2018). Second, control users in the SMHD are not mapped 1:1 to diagnosed users, but only to a specific disorder, e.g. controls for bipolar disorder. Using a balanced dataset would have allowed for a closer comparison of the present work to the work of Sekulić et al. (2018). Due to the lack of such mapping, such comparison was not possible. Third, according to Yahav et al. (2018), the use of TF-IDF for analysis of internet comments can introduce a bias since the frequencies of terms used in online discourses may be inflated due to their dependent nature. In other words, the frequency of a particular term can be inflated by the fact that the participants are sharing a discussion on a given topic. This is particularly an issue in threaded forums such as Reddit where users reply to each other's comments, creating repetition of terms in the discussion. Since comment-thread information was not preserved in the dataset, the possibility of this issue remains.

Furthermore, the present work is limited to the use of relatively simple ML algorithms and TF-IDF-weighted bag of words which is based on a unigram approach. It is possible that more complex ML algorithms and NLP techniques such as neural networks and n-gram models may perform better at distinguishing mental health disorders. However, the primary purpose of this study was to test the techniques which previous studies had introduced. Additionally, SVM and RF algorithms for TF-IDF predictors had to be omitted from the analysis due to the high computational requirements of training and testing the 160,000 predictor models. Finally, there is a question of external validity of this and similar studies. The study design relies on the identification of users who have already been diagnosed with bipolar disorder, while the target group would be users who have not yet been diagnosed. A more robust approach would perhaps rely on social media activity of a user over years prior to their diagnosis.

The main contribution of this study is that it highlights the importance of adding robustness checks to studies predicting specific mental health disorders from social media. Future research should replicate the findings for other disorders and across different social media platforms involving more complex ML algorithms and NLP models. Overall, further research is needed to establish how different mental health conditions affect social media activity and writing styles. In conclusion, the findings of this study seem to indicate that we may be able to detect mental health problems in general from user activity but perhaps more work is needed to be able to detect specific mental health disorders.

References

Amir, S., Dredze, M., & Ayers, J. W. (2019, June). Mental health surveillance over social media with digital cohorts. In Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology (pp. 114-120).

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders: DSM-5* (5th ed). American Psychiatric Association. . https://doi.org/10.1176/appi.books.9780890425596

Angst, J., Azorin, J. M., Bowden, C. L., Perugi, G., Vieta, E., Gamma, A., Young, A. H., & BRIDGE Study Group. (2011). Prevalence and characteristics of undiagnosed bipolar disorders in patients with a major depressive episode: the BRIDGE study. Archives of general psychiatry, 68(8), 791-799. https://doi.org/10.1001/archgenpsychiatry.2011.87

Baldessarini, R. J., Tondo, L., Baethge, C. J., Lepri, B., & Bratti, I. M. (2007). Effects of treatment latency on response to maintenance treatment in manic-depressive disorders. *Bipolar Disorders*, *9*(4), 386–393. https://doi.org/10.1111/j.1399-5618.2007.00385.x

Caspi, A., Houts, R. M., Ambler, A., Danese, A., Elliott, M. L., Hariri, A., Harrington, H., Hogan, S., Poulton, R., Ramrakha, S., Rasmussen, L. J. H., Reuben, A., Richmond-Rakerd, L., Sugden, K., Wertz, J., Williams, B. S., & Moffitt, T. E. (2020). Longitudinal Assessment of Mental Health Disorders and Comorbidities Across 4 Decades Among Participants in the Dunedin Birth Cohort Study. https://doi.org/10.1001/jamanetworkopen.2020.3221

Cohan, A., Desmet, B., Yates, A., Soldaini, L., MacAvaney, S., & Goharian, N. (2018). SMHD: a large-scale resource for exploring online language usage for multiple mental health conditions. arXiv preprint arXiv:1806.05258.

Coppersmith, G., Harman, C., & Dredze, M. (2014). Measuring post traumatic stress disorder in Twitter. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 8, No. 1).

De Choudhury, M., Counts, S., & Horvitz, E. (2013). Social media as a measurement tool of depression in populations. In Proceedings of the 5th annual ACM web science conference (pp. 47-56).

Hirschfeld, R. M., Lewis, L., & Vornik, L. A. (2003). Perceptions and impact of bipolar disorder: how far have we really come? Results of the national depressive and manic-depressive association 2000 survey of individuals with bipolar disorder. The Journal of clinical psychiatry, 64(2), 161–174.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning: with Applications in R (Springer Texts in Statistics) (1st ed. 2013 ed.). Springer.

Jiang, Z. P., Levitan, S. I., Zomick, J., & Hirschberg, J. (2020, November). Detection of mental health from Reddit via deep contextualized representations. In Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis (pp. 147-156).

Jurafsky, D., & Martin, J. M. (2021). Speech and Language Processing An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition (Third Edition). Random House.

Kessler, R. C., Berglund, P., Demler, O., Jin, R., Merikangas, K. R., & Walters, E. E. (2005). Lifetime Prevalence and Age-of-Onset Distributions of DSM-IV Disorders in the National Comorbidity Survey Replication. *Archives of General Psychiatry*, *62*(6), 593–602. https://doi.org/10.1001/archpsyc.62.6.593

Lish, J. D., Dime-Meenan, S., Whybrow, P. C., Price, R. A., & Hirschfeld, R. M. A. (1994). The National Depressive and Manic-depressive Association (DMDA) survey of bipolar members. *Journal of Affective Disorders*, *31*(4), 281–294. https://doi.org/10.1016/0165-0327(94)90104-X

Mitchell, M., Hollingshead, K., & Coppersmith, G. (2015). Quantifying the language of schizophrenia in social media. In Proceedings of the 2nd workshop on Computational linguistics and clinical psychology: From linguistic signal to clinical reality (pp. 11-20).

Müller, A. C., & Guido, S. (2016). Introduction to machine learning with Python: a guide for data scientists. O'Reilly Media, Inc.

Noble, W. S. (2006). What is a support vector machine?. Nature biotechnology, 24(12), 1565-1567.

OECD & European Union (2018), Health at a Glance: Europe 2018: State of Health in the EU Cycle, OECD Publishing, https://doi.org/10.1787/health_glance_eur-2018-en

Patel, V., Flisher, A. J., Hetrick, S., & McGorry, P. (2007). Mental health of young people: A global public-health challenge. The Lancet, 369(9569), 1302–1313. https://doi.org/10.1016/S0140-6736(07)60368-7

Paykel, E. S., Tylee, A., Wright, A., Priest, R. G., & et al. (1997). The Defeat Depression Campaign: Psychiatry in the public arena. *The American Journal of Psychiatry*, *154*(6, Suppl), 59–65. https://doi.org/10.1176/ajp.154.6.59

Peng, C. Y. J., Lee, K. L., & Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. The journal of educational research, 96(1), 3-14.

Pennebaker, J.W., Boyd, R.L., Jordan, K., & Blackburn, K. (2015a). The development and psychometric properties of LIWC2015. Austin, TX: University of Texas at Austin

Pennebaker, J. W., Booth, R. J, Boyd, R. L. &. Francis, M. E. (2015b). Linguistic Inquiry and Word Count: LIWC 2015.

Perlis, R. H. (2005). Misdiagnosis of bipolar disorder. The American journal of managed care, 11(9 Suppl), S271-4.

Phillips, M. L., & Kupfer, D. J. (2013). Bipolar disorder diagnosis: challenges and future directions. The Lancet, 381(9878), 1663-1671

Rajaraman, A., & Ullman, J. D. (2011). Mining of Massive Datasets (1st ed.). Cambridge University Press.

Ramos, J. (2003, December). Using tf-idf to determine word relevance in document queries. In Proceedings of the first instructional conference on machine learning (Vol. 242, No. 1, pp. 29-48).

Ritchie, H., & Roser, M. (2018). Mental Health. *Our World in Data*. https://ourworldindata.org/mental-health

Sekulić, I., Gjurković, M., & Šnajder, J. (2018). Not just depressed: Bipolar disorder prediction on reddit. arXiv preprint arXiv:1811.04655.

Shen, J. H., & Rudzicz, F. (2017, August). Detecting anxiety through reddit. In Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology—From Linguistic Signal to Clinical Reality (pp. 58-65).

UN Secretary-General Policy. (2020). *COVID-19 and Need for Action on Mental Health*.

Williams, S. Z., Chung, G. S., & Muennig, P. A. (2017). Undiagnosed depression: A community diagnosis. *SSM - Population Health*, *3*, 633–638. https://doi.org/10.1016/j.ssmph.2017.07.012

Yahav, I., Shehory, O., & Schwartz, D. (2018). Comments mining with TF-IDF: the inherent bias and its removal. IEEE Transactions on Knowledge and Data Engineering, 31(3), 437-450.

Zheng, A., & Casari, A. (2018). Feature engineering for machine learning: principles and techniques for data scientists. " O'Reilly Media, Inc.".

Appendix A: Descriptives of LIWC Variables with Means and Standard Deviations

	Bipolar mean			Bipolar sd			Autism mean			Autism sd			Depression mean			Depression sd			Healthy Control Bipolar mean			Healthy Control Bipolar sd		
	dev	test	train	dev	test	train	dev	test	train	dev	test	train	dev	test	train	dev	test	train	dev	test	train	dev	test	train
WC	6965	6936	7421	7019	7266	7806	8057	7724	7525	6324	5953	8014	7196	7165	7343	6799	6907	5985	7922	7821	8011	6910	6944	6628
Analytic	45.3	45.7	46.2	13.8	13.6	14.0	49.5	48.3	49.1	14.8	14.9	14.9	45.8	45.4	45.3	14.2	14.2	14.4	56.6	56.8	57.0	14.3	14.6	14.5
Clout	47.9	48.8	48.3	12.9	12.5	12.9	49.0	46.9	47.6	13.1	12.0	12.8	47.9	48.6	48.2	12.7	13.1	13.3	52.5	52.9	52.7	11.5	11.4	11.4
Authentic	53.2	52.3	53.0	16.4	17.0	16.7	47.1	47.5	48.1	17.5	18.1	18.8	52.2	51.9	52.2	16.9	17.1	17.4	37.9	37.7	37.6	16.7	16.6	16.7
Tone	54.2	53.9	54.3	20.8	20.3	20.1	48.3	49.1	49.2	18.8	20.3	20.1	53.9	54.2	54.1	19.9	20.6	20.0	58.2	57.8	57.9	20.6	20.5	20.3
WPS	21.4	22.3	21.9	10.1	25.4	13.9	33.3	28.9	33.1	117.6	28.2	72.4	24.2	23.3	23.2	31.6	21.0	14.7	45.2	46.4	45.0	328.5	359.1	292.2
Sixltr	14.8	14.8	14.8	2.2	2.2	2.0	15.6	15.5	15.4	2.6	2.3	2.5	14.8	14.8	14.8	2.1	2.1	2.1	14.8	14.9	14.9	3.1	3.3	3.3
Dic	85.0	85.1	85.2	4.9	5.5	4.7	83.2	82.8	82.9	5.8	6.8	6.5	84.6	84.7	84.8	5.6	5.7	5.4	79.2	79.2	79.0	9.4	9.7	9.7
function	52.7	52.6	52.6	4.1	4.5	4.0	51.7	51.5	51.5	4.6	5.3	5.1	52.4	52.4	52.5	4.5	4.5	4.4	47.8	47.8	47.7	7.4	7.6	7.7
pronoun	16.6	16.5	16.5	2.4	2.5	2.5	15.5	15.6	15.5	2.7	2.7	2.7	16.4	16.4	16.5	2.6	2.6	2.6	14.1	14.1	14.1	3.1	3.2	3.2
ppron	10.7	10.7	10.7	2.1	2.2	2.2	9.7	9.7	9.6	2.3	2.3	2.3	10.5	10.6	10.6	2.2	2.2	2.2	8.7	8.6	8.6	2.3	2.3	2.3
i	6.2	6.0	6.1	1.8	1.8	1.8	5.3	5.5	5.4	1.9	1.9	2.0	6.0	6.0	6.0	1.8	1.8	1.9	4.4	4.4	4.4	1.7	1.7	1.7
we	0.5	0.5	0.5	0.3	0.3	0.3	0.4	0.4	0.5	0.3	0.3	0.4	0.5	0.5	0.5	0.3	0.3	0.3	0.5	0.5	0.4	0.4	0.3	0.3
you	2.0	2.0	2.0	0.8	0.9	0.9	1.9	1.8	1.8	0.9	0.8	0.8	1.9	2.0	2.0	0.9	0.9	0.9	2.1	2.1	2.1	1.0	1.0	1.0
shehe	1.3	1.4	1.3	0.9	0.9	1.0	1.1	1.1	1.1	0.8	0.9	0.8	1.3	1.3	1.3	0.9	0.9	1.0	1.0	1.0	1.0	0.7	0.7	0.7
they	0.8	0.8	0.8	0.3	0.4	0.3	0.9	0.8	0.8	0.4	0.4	0.4	0.8	0.8	0.8	0.4	0.4	0.4	0.8	0.8	0.8	0.4	0.4	0.4
ipron	5.8	5.8	5.8	1.0	1.0	0.9	5.9	5.9	5.8	1.0	1.0	1.0	5.8	5.8	5.8	1.0	1.0	1.0	5.5	5.5	5.4	1.3	1.3	1.3
article	5.7	5.7	5.8	0.9	0.9	0.9	5.9	5.8	5.9	1.0	1.0	1.0	5.7	5.7	5.7	0.9	0.9	0.9	6.0	6.0	6.0	1.3	1.3	1.3
prep	11.9	11.9	12.0	1.1	1.2	1.2	11.9	11.7	11.8	1.3	1.4	1.4	11.9	11.8	11.9	1.2	1.2	1.3	11.0	11.0	11.0	1.8	1.8	1.9
auxverb	9.5	9.5	9.4	1.2	1.2	1.2	9.4	9.4	9.4	1.3	1.3	1.3	9.4	9.4	9.5	1.2	1.2	1.2	8.7	8.7	8.6	1.7	1.7	1.8
adverb	5.7	5.7	5.7	0.9	0.9	0.9	5.6	5.7	5.7	1.0	1.0	1.0	5.8	5.8	5.8	0.9	0.9	1.0	5.2	5.2	5.1	1.1	1.2	1.2
conj	6.5	6.5	6.5	1.0	1.0	1.0	6.4	6.4	6.4	1.0	1.1	1.1	6.5	6.5	6.5	1.0	1.0	1.0	5.5	5.5	5.5	1.2	1.2	1.3
negate	2.0	2.0	2.0	0.5	0.5	0.5	2.1	2.1	2.1	0.5	0.6	0.6	2.0	2.0	2.0	0.5	0.5	0.5	1.9	1.9	1.9	0.6	0.6	0.7
verb	17.6	17.5	17.5	1.8	2.0	1.8	17.0	17.0	17.0	2.1	2.3	2.2	17.5	17.6	17.6	2.0	2.0	2.0	16.0	16.0	15.9	2.8	2.8	2.9
adj	4.8	4.8	4.8	0.7	0.7	0.7	4.8	4.7	4.8	0.7	0.7	0.7	4.8	4.8	4.8	0.7	0.7	0.7	4.7	4.7	4.7	1.2	1.1	1.1
compare	2.4	2.4	2.4	0.5	0.5	0.4	2.5	2.5	2.5	0.5	0.5	0.5	2.5	2.5	2.5	0.5	0.5	0.5	2.3	2.3	2.3	0.6	0.6	0.6
interrog	1.5	1.5	1.5	0.4	0.4	0.4	1.6	1.6	1.5	0.4	0.4	0.4	1.5	1.5	1.5	0.4	0.4	0.4	1.4	1.4	1.4	0.5	0.5	0.9
number	1.7	1.7	1.7	0.8	0.7	0.6	1.7	1.8	1.8	0.9	1.2	0.8	1.8	1.7	1.7	0.8	0.8	0.7	2.3	2.3	2.3	1.6	1.7	1.4
quant	2.3	2.3	2.3	0.4	0.5	0.4	2.4	2.3	2.3	0.4	0.6	0.5	2.3	2.3	2.3	0.4	0.4	0.4	2.2	2.2	2.2	0.9	0.8	1.1
affect	6.2	6.2	6.1	1.3	1.3	1.3	5.8	5.8	5.8	1.3	1.3	1.2	6.1	6.1	6.1	1.2	1.3	1.3	6.0	6.0	6.0	1.9	1.8	1.9
posemo	3.8	3.8	3.8	1.1	1.1	1.1	3.5	3.5	3.5	1.0	1.0	1.0	3.8	3.8	3.8	1.0	1.1	1.1	3.9	3.9	3.9	1.6	1.6	1.6

negemo	2.3	2.3	2.2	0.7	0.7	0.7	2.2	2.2	2.2	0.7	0.8	0.7	2.2	2.2	2.2	0.7	0.7	0.7	2.0	2.0	2.0	0.9	0.8	0.9
anx	0.3	0.3	0.3	0.2	0.1	0.2	0.3	0.3	0.3	0.2	0.2	0.2	0.3	0.3	0.3	0.2	0.2	0.2	0.2	0.2	0.2	0.1	0.1	0.1
anger	0.8	0.8	0.8	0.4	0.4	0.4	0.8	0.8	0.8	0.4	0.6	0.5	0.8	0.8	0.8	0.4	0.4	0.4	0.9	0.9	0.9	0.7	0.6	0.7
sad	0.4	0.4	0.4	0.2	0.2	0.2	0.4	0.4	0.4	0.2	0.1	0.2	0.4	0.4	0.4	0.2	0.2	0.2	0.3	0.3	0.3	0.2	0.2	0.2
social	9.4	9.5	9.4	2.2	2.2	2.2	9.1	8.9	8.9	2.3	2.2	2.2	9.3	9.4	9.4	2.3	2.4	2.4	8.4	8.4	8.4	2.2	2.2	2.2
family	0.5	0.5	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.3	0.3	0.3	0.3	0.5	0.3
friend	0.4	0.4	0.4	0.2	0.2	0.2	0.4	0.4	0.4	0.3	0.3	0.2	0.4	0.4	0.4	0.2	0.3	0.2	0.4	0.4	0.4	0.2	0.2	0.3
female	0.9	0.9	0.9	0.7	0.7	0.7	0.8	0.7	0.7	0.7	0.7	0.7	0.9	0.9	0.9	0.7	0.7	0.8	0.5	0.5	0.5	0.6	0.6	0.7
male	1.2	1.2	1.2	0.7	0.8	0.8	1.0	1.1	1.1	0.6	0.7	0.7	1.2	1.2	1.2	0.8	0.8	1.1	1.1	1.1	1.1	0.7	0.7	0.7
cogproc	13.0	13.0	12.9	1.9	1.9	1.8	13.3	13.3	13.3	2.0	2.1	2.2	13.1	13.1	13.1	1.9	1.9	1.9	11.7	11.7	11.7	2.3	2.4	2.4
insight	2.5	2.5	2.4	0.6	0.6	0.6	2.5	2.5	2.5	0.6	0.7	0.6	2.5	2.5	2.5	0.6	0.6	0.6	2.1	2.1	2.1	0.6	0.6	0.6
cause	1.7	1.7	1.7	0.4	0.4	0.4	1.8	1.9	1.9	0.4	0.4	0.5	1.7	1.8	1.7	0.4	0.4	0.4	1.7	1.7	1.6	0.5	0.5	0.5
discrep	1.8	1.9	1.8	0.4	0.4	0.4	1.9	1.9	1.9	0.5	0.5	0.5	1.9	1.9	1.9	0.4	0.4	0.4	1.8	1.8	1.8	0.5	0.5	0.5
tentat	3.3	3.4	3.3	0.7	0.7	0.7	3.5	3.4	3.5	0.8	0.7	0.8	3.4	3.4	3.4	0.7	0.7	0.7	3.1	3.1	3.1	0.8	0.8	0.8
certain	1.6	1.6	1.6	0.4	0.4	0.4	1.5	1.5	1.5	0.4	0.4	0.4	1.6	1.6	1.6	0.4	0.4	0.4	1.5	1.5	1.4	0.4	0.4	0.4
differ	3.8	3.8	3.8	0.8	0.7	0.7	3.9	3.9	3.9	0.8	0.8	0.9	3.8	3.8	3.8	0.8	0.8	0.8	3.4	3.4	3.4	0.9	0.9	0.9
percept	2.7	2.7	2.7	0.6	0.7	0.6	2.5	2.5	2.5	0.7	0.6	0.6	2.7	2.7	2.7	0.6	0.7	0.7	2.7	2.7	2.7	1.0	0.9	1.0
see	1.1	1.1	1.1	0.4	0.5	0.4	1.1	1.1	1.1	0.4	0.4	0.4	1.1	1.1	1.1	0.4	0.4	0.5	1.3	1.3	1.3	0.6	0.6	0.7
hear	0.7	0.7	0.7	0.3	0.3	0.3	0.7	0.7	0.7	0.3	0.3	0.3	0.7	0.7	0.7	0.3	0.3	0.3	0.6	0.6	0.6	0.4	0.4	0.4
feel	0.7	0.7	0.7	0.3	0.3	0.3	0.6	0.6	0.6	0.3	0.3	0.3	0.7	0.7	0.7	0.3	0.3	0.3	0.6	0.5	0.6	0.5	0.3	0.3
bio	2.6	2.6	2.6	1.1	1.0	1.2	2.2	2.1	2.1	1.0	1.1	0.9	2.5	2.5	2.5	1.1	1.0	1.0	2.1	2.1	2.1	1.4	1.4	1.4
body	0.8	0.8	0.8	0.4	0.4	0.8	0.7	0.7	0.7	0.4	0.5	0.4	0.8	0.8	0.8	0.4	0.4	0.4	0.8	0.8	0.8	0.7	0.7	0.7
health	0.8	0.8	0.8	0.5	0.5	0.4	0.7	0.6	0.6	0.5	0.4	0.4	0.7	0.7	0.7	0.5	0.4	0.4	0.4	0.4	0.4	0.3	0.3	0.3
sexual	0.3	0.4	0.4	0.3	0.3	0.4	0.3	0.3	0.3	0.3	0.4	0.3	0.3	0.3	0.3	0.3	0.4	0.3	0.3	0.3	0.3	0.7	0.6	0.7
ingest	0.7	0.6	0.6	0.6	0.5	0.5	0.5	0.4	0.5	0.5	0.4	0.4	0.6	0.6	0.6	0.5	0.5	0.5	0.5	0.5	0.5	0.6	0.5	0.5
drives	6.7	6.7	6.7	1.0	1.0	0.9	6.5	6.5	6.5	1.0	1.1	1.1	6.6	6.7	6.7	1.0	1.0	1.0	6.5	6.5	6.5	1.5	1.5	1.6
affiliation	1.8	1.9	1.8	0.7	0.7	0.6	1.7	1.7	1.8	0.7	0.7	0.7	1.8	1.9	1.9	0.7	0.7	0.7	1.7	1.7	1.7	0.7	0.8	0.7
achieve	1.4	1.4	1.4	0.4	0.4	0.4	1.3	1.4	1.4	0.4	0.4	0.5	1.4	1.4	1.4	0.4	0.4	0.4	1.4	1.4	1.4	0.5	0.5	0.6
power	2.1	2.1	2.1	0.5	0.5	0.4	2.1	2.1	2.1	0.5	0.6	0.5	2.1	2.1	2.1	0.5	0.5	0.5	2.1	2.1	2.1	0.7	0.7	0.7
reward	1.6	1.6	1.6	0.4	0.5	0.4	1.4	1.5	1.5	0.4	0.6	0.4	1.6	1.6	1.6	0.4	0.4	0.4	1.6	1.6	1.6	0.9	0.8	1.0
risk	0.5	0.5	0.5	0.2	0.2	0.2	0.6	0.5	0.5	0.2	0.2	0.2	0.5	0.5	0.5	0.2	0.2	0.2	0.5	0.5	0.5	0.2	0.3	0.2
focuspast	3.9	3.9	3.9	1.1	1.2	1.1	3.5	3.6	3.6	1.0	1.2	1.2	3.9	3.8	3.9	1.1	1.1	1.1	3.3	3.2	3.3	1.3	1.2	1.3
focuspresent	11.5	11.5	11.4	1.6	1.7	1.6	11.4	11.4	11.3	1.7	1.8	1.7	11.5	11.6	11.6	1.7	1.7	1.7	10.8	10.8	10.8	2.0	2.0	2.1
focusfuture	1.2	1.2	1.1	0.3	0.3	0.3	1.1	1.1	1.1	0.3	0.3	0.3	1.1	1.2	1.2	0.3	0.3	0.3	1.1	1.1	1.1	0.4	0.4	0.4
relativ	12.3	12.3	12.4	1.6	1.8	1.7	11.7	11.6	11.8	1.7	1.9	1.8	12.2	12.2	12.3	1.7	1.7	1.8	11.7	11.7	11.7	2.1	2.2	2.2
motion	1.7	1.7	1.7	0.4	0.4	0.4	1.6	1.6	1.6	0.4	0.4	0.4	1.7	1.7	1.7	0.4	0.4	0.4	1.7	1.7	1.7	0.5	0.5	0.7

space	6.0	6.0	6.1	0.8	0.9	0.9	5.8	5.7	5.9	0.9	1.0	1.0	5.9	5.9	6.0	0.9	0.9	0.9	5.9	5.9	5.9	1.2	1.2	1.2
time	4.7	4.7	4.8	1.0	1.0	1.0	4.4	4.4	4.4	1.0	1.0	1.0	4.7	4.7	4.8	1.0	1.0	1.0	4.3	4.3	4.3	1.0	1.0	1.1
work	1.8	1.8	1.8	0.8	0.9	0.9	1.8	1.8	1.8	0.9	0.9	0.8	1.8	1.8	1.8	0.9	0.8	0.9	1.8	1.8	1.8	1.1	1.1	1.0
leisure	1.2	1.2	1.3	0.6	0.6	0.6	1.3	1.3	1.4	0.7	0.8	0.8	1.2	1.3	1.2	0.6	0.6	0.6	1.6	1.6	1.6	0.9	0.9	0.9
home	0.4	0.4	0.4	0.3	0.2	0.2	0.3	0.3	0.3	0.2	0.2	0.2	0.4	0.4	0.4	0.3	0.3	0.3	0.3	0.3	0.3	0.2	0.4	0.4
money	0.7	0.7	0.7	0.5	0.5	0.4	0.6	0.6	0.6	0.4	0.5	0.4	0.7	0.7	0.7	0.5	0.5	0.4	0.9	0.9	0.9	0.8	0.8	0.8
relig	0.3	0.3	0.3	0.4	0.4	0.3	0.3	0.3	0.2	0.4	0.4	0.3	0.3	0.3	0.2	0.3	0.3	0.3	0.3	0.3	0.3	0.4	0.4	0.4
death	0.2	0.2	0.2	0.1	0.2	0.1	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.1	0.2	0.1	0.2	0.2	0.2	0.3	0.2	0.2
informal	1.9	1.9	1.9	1.0	1.0	1.1	1.8	1.8	1.8	1.0	1.0	0.9	1.9	1.9	1.9	0.9	1.2	0.9	2.5	2.5	2.5	1.6	1.5	2.1
swear	0.5	0.5	0.5	0.4	0.4	0.8	0.4	0.4	0.4	0.4	0.5	0.4	0.4	0.4	0.4	0.3	0.4	0.4	0.6	0.6	0.6	0.8	0.7	0.8
netspeak	0.8	0.8	0.8	0.6	0.6	0.6	0.8	0.8	0.8	0.7	0.6	0.6	0.8	0.8	0.8	0.6	0.7	0.6	1.3	1.3	1.3	1.2	1.1	1.4
assent	0.3	0.3	0.3	0.2	0.2	0.2	0.3	0.3	0.3	0.2	0.2	0.2	0.3	0.3	0.3	0.2	0.7	0.2	0.4	0.4	0.4	0.3	0.3	0.9
nonflu	0.2	0.2	0.2	0.1	0.1	0.1	0.2	0.2	0.2	0.1	0.1	0.1	0.2	0.2	0.2	0.1	0.1	0.1	0.3	0.3	0.3	0.2	0.3	0.8
filler	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
AllPunc	32.4	32.3	31.9	8.5	8.6	8.5	32.9	33.6	33.3	8.8	10.3	9.6	32.6	32.7	32.3	8.9	9.0	8.6	42.9	47.1	43.3	18.0	469.0	19.9
Period	6.9	7.0	7.0	2.0	2.0	2.0	6.5	6.3	6.3	1.9	1.9	1.9	6.6	6.7	6.6	2.0	2.0	2.0	7.2	7.2	7.2	3.2	3.4	3.3
Comma	6.7	6.6	6.6	2.2	2.1	2.1	6.9	7.0	6.9	2.2	2.5	2.2	6.8	6.7	6.7	2.2	2.2	2.2	8.8	9.9	9.0	4.4	116.0	5.2
Colon	0.2	0.2	0.2	0.2	0.2	0.2	0.3	0.3	0.3	0.4	0.6	0.3	0.3	0.3	0.3	0.3	0.4	0.3	0.4	0.4	0.4	0.6	0.6	0.6
SemiC	0.2	0.2	0.2	0.2	0.3	0.2	0.2	0.2	0.2	0.2	0.3	0.3	0.2	0.2	0.2	0.4	0.2	0.3	0.2	0.2	0.2	0.4	0.4	0.8
QMark	0.8	0.8	0.8	0.5	0.5	0.5	0.8	0.9	0.8	0.6	0.7	0.6	0.8	0.8	0.8	0.9	0.5	0.5	1.1	1.1	1.1	0.7	0.8	0.7
Exclam	0.9	0.9	0.8	1.1	1.0	0.9	0.6	0.6	0.6	0.6	0.9	0.7	0.8	0.9	0.8	0.9	0.9	0.9	1.0	1.0	1.0	1.5	1.5	2.0
Dash	0.7	0.7	0.7	0.8	0.7	0.6	0.8	0.8	0.7	0.7	0.9	0.7	0.7	0.7	0.7	0.7	0.8	0.6	1.1	1.0	1.0	4.5	1.5	2.0
Quote	7.3	7.3	7.2	4.2	4.0	4.0	7.2	7.4	7.4	4.2	5.0	4.3	7.3	7.4	7.2	4.1	4.2	3.9	12.3	14.5	12.6	9.0	233.7	10.7
Apostro	3.1	3.1	3.1	1.1	1.0	1.1	3.0	3.1	3.1	1.2	1.3	1.2	3.1	3.1	3.1	1.1	1.1	1.1	2.6	2.6	2.6	1.1	1.1	1.1
Parenth	0.8	0.9	0.8	0.8	0.7	0.6	1.0	1.1	1.0	0.8	1.4	0.7	0.9	0.9	0.9	0.8	0.7	0.7	1.1	1.0	1.0	1.4	1.2	1.3
OtherP	4.8	4.7	4.6	3.1	3.4	3.1	5.6	5.8	5.9	4.1	3.7	4.7	5.1	5.1	5.0	3.5	4.1	3.7	7.1	8.1	7.1	8.2	119.6	7.9

Appendix B: Parameter Grids

Logistic Regression

```
param_grid = {
            'penalty' : ['l1', 'l2'],
          'solver' : [ 'liblinear']
       }
```

Random Forest

```
  param_grid = {
          'n_estimators': [300, 1000, 1500],
          'max_depth': [10,30,60,None],
          'max_features': ['auto','sqrt'],
          'bootstrap': [True, False],
          'min_samples_split' : [2, 5, 10]
       }
```

Support Vector Machine

```
param_grid = {
          'C': [0.1,1, 10, 60,100],
          'gamma': [0.1, 0.01, 0.001],
          'kernel': ['linear','sigmoid']
       }
```

Appendix C: Robustness Checks

A number of robustness checks were performed based on logistic regression. They are (1) running the analysis which also includes healthy controls for depression and autism in addition to the bipolar healthy controls, (2) exclusion of all healthy controls, (3) exclusion of all comorbid disorders, and (4) hyperparameter tuning based on recall instead of F1 value.

1. **Including healthy controls for depression and autism – 3rd Specification**

User behavior only

Best params:
{'penalty': 'l1', 'solver': 'liblinear'}
0.012959251250364045
Predicting test outcome:
Classification report:

	precision	recall	f1-score	support
0	0.97	1.00	0.99	43018
1	0.37	0.01	0.02	1247
accuracy			0.97	44265
macro avg	0.67	0.51	0.51	44265
weighted avg	0.96	0.97	0.96	44265

Confusion matrix:
[[16 0 0 1231]
[11 0 0 2260]
[5 0 0 492]
[11 0 0 40239]]

User behavior + LIWC

Best params:
{'penalty': 'l1', 'solver': 'liblinear'}
0.06974739833618529
Predicting test outcome:
Classification report:

	precision	recall	f1-score	support
0	0.97	1.00	0.99	43018
1	0.39	0.04	0.07	1247
accuracy			0.97	44265
macro avg	0.68	0.52	0.53	44265
weighted avg	0.96	0.97	0.96	44265

Confusion matrix:
[[51 0 0 1196]
[61 0 0 2210]
[6 0 0 491]
[14 0 0 40236]]

LIWC only

Best params:
{'penalty': 'l1', 'solver': 'liblinear'}
0.06926843657728021
Predicting test outcome:
Classification report:

	precision	recall	f1-score	support
0	0.97	1.00	0.99	43018
1	0.34	0.03	0.06	1247
accuracy			0.97	44265
macro avg	0.65	0.52	0.52	44265
weighted avg	0.95	0.97	0.96	44265

Confusion matrix:
[[40 0 0 1207]
[56 0 0 2215]
[7 0 0 490]
[16 0 0 40234]]

2. No healthy controls – 3rd Specification

User behavior only

Best params:
{'penalty': 'l1', 'solver': 'liblinear'}
0.002493077181873862
Predicting test outcome:
Classification report:

	precision	recall	f1-score	support
0	0.69	1.00	0.82	2768
1	0.00	0.00	0.00	1247
accuracy			0.69	4015
macro avg	0.34	0.50	0.41	4015
weighted avg	0.48	0.69	0.56	4015

Confusion matrix:
[[0 0 0 1247]
[0 0 0 2271]
[0 0 0 497]
[0 0 0 0]]

User behavior + LIWC

Best params:
{'penalty': 'l1', 'solver': 'liblinear'}

0.10392609691303958
Predicting test outcome:
Classification report:

	precision	recall	f1-score	support
0	0.70	0.98	0.81	2768
1	0.48	0.05	0.09	1247
accuracy			0.69	4015
macro avg	0.59	0.51	0.45	4015
weighted avg	0.63	0.69	0.59	4015

Confusion matrix:
[[63 0 0 1184]
[62 0 0 2209]
[6 0 0 491]
[0 0 0 0]]

LIWC only

Best params:
{'penalty': 'l1', 'solver': 'liblinear'}
0.10548566368902627
Predicting test outcome:
Classification report:

	precision	recall	f1-score	support
0	0.70	0.97	0.81	2768
1	0.48	0.06	0.11	1247
accuracy			0.69	4015
macro avg	0.59	0.52	0.46	4015
weighted avg	0.63	0.69	0.59	4015

Confusion matrix:
[[76 0 0 1171]
[77 0 0 2194]
[5 0 0 492]
[0 0 0 0]]

3. No comorbidities – 1st, 2nd and 3rd Specification

User behavior only

1st Specification:

Best params:
{'penalty': 'l1', 'solver': 'liblinear'}
0.14506689837793613
Predicting test outcome:
Classification report:

	precision	recall	f1-score	support
0	0.96	1.00	0.98	14368
1	0.77	0.11	0.19	735
accuracy			0.96	15103
macro avg	0.86	0.55	0.58	15103
weighted avg	0.95	0.96	0.94	15103

Confusion matrix:
```
[[   80     0     0   655]
 [    0     0     0     0]
 [    0     0     0     0]
 [   24     0     0 14344]]
```

2nd Specification:

Classification report:

	precision	recall	f1-score	support
0	0.96	0.99	0.98	16045
1	0.33	0.11	0.16	735
accuracy			0.95	16780
macro avg	0.65	0.55	0.57	16780
weighted avg	0.93	0.95	0.94	16780

confusion matrix:
```
[[   80     0     0   655]
 [  110     0     0  1206]
 [   25     0     0   336]
 [   24     0     0 14344]]
```

3rd Specification:

Best params:

{'penalty': 'l1', 'solver': 'liblinear'}
0.01712335833125126
Predicting test outcome:
Classification report:

	precision	recall	f1-score	support
0	0.96	1.00	0.98	16045
1	0.43	0.01	0.03	735
accuracy			0.96	16780
macro avg	0.70	0.51	0.50	16780
weighted avg	0.93	0.96	0.94	16780

Confusion matrix:
```
[[   10     0     0   725]
```

[7 0 0 1309]
[2 0 0 359]
[4 0 0 14364]]

User behavior + LIWC

1st Specification

Best params:
{'penalty': 'l1', 'solver': 'liblinear'}
0.40710757225998434
Predicting test outcome:
Classification report:

	precision	recall	f1-score	support
0	0.97	0.99	0.98	14368
1	0.71	0.30	0.42	735
accuracy			0.96	15103
macro avg	0.84	0.65	0.70	15103
weighted avg	0.95	0.96	0.95	15103

Confusion matrix:
[[222 0 0 513]
[0 0 0 0]
[0 0 0 0]
[92 0 0 14276]]

2nd Specification

Classification report:

	precision	recall	f1-score	support
0	0.97	0.97	0.97	16045
1	0.31	0.30	0.31	735
accuracy			0.94	16780
macro avg	0.64	0.64	0.64	16780
weighted avg	0.94	0.94	0.94	16780

confusion matrix:
[[222 0 0 513]
[330 0 0 986]
[69 0 0 292]
[92 0 0 14276]]

3rd Specification

Best params:
{'penalty': 'l1', 'solver': 'liblinear'}
0.07883293061787867

Predicting test outcome:
Classification report:

	precision	recall	f1-score	support
0	0.96	1.00	0.98	16045
1	0.37	0.05	0.08	735
accuracy			0.95	16780
macro avg	0.66	0.52	0.53	16780
weighted avg	0.93	0.95	0.94	16780

Confusion matrix:
[[34 0 0 701]
[48 0 0 1268]
[3 0 0 358]
[8 0 0 14360]]

LIWC only

1st Specification

Best params:
{'penalty': 'l1', 'solver': 'liblinear'}
0.35340809462325734
Predicting test outcome:
Classification report:

	precision	recall	f1-score	support
0	0.96	0.99	0.98	14368
1	0.68	0.25	0.37	735
accuracy			0.96	15103
macro avg	0.82	0.62	0.67	15103
weighted avg	0.95	0.96	0.95	15103

Confusion matrix:
[[185 0 0 550]
[0 0 0 0]
[0 0 0 0]
[86 0 0 14282]]

2nd Specification

Classification report:

	precision	recall	f1-score	support
0	0.97	0.97	0.97	16045
1	0.30	0.25	0.28	735
accuracy			0.94	16780
macro avg	0.64	0.61	0.62	16780

weighted avg 0.94 0.94 0.94 16780

confusion matrix:
[[185 0 0 550]
[274 0 0 1042]
[63 0 0 298]
[86 0 0 14282]]

3rd Specification

Best params:
{'penalty': 'l1', 'solver': 'liblinear'}
0.07105527274765279
Predicting test outcome:
Classification report:

	precision	recall	f1-score	support
0	0.96	1.00	0.98	16045
1	0.34	0.04	0.07	735
accuracy			0.95	16780
macro avg	0.65	0.52	0.52	16780
weighted avg	0.93	0.95	0.94	16780

Confusion matrix:
[[27 0 0 708]
[39 0 0 1277]
[5 0 0 356]
[8 0 0 14360]]

4. Hyperparameter tuning based on recall (instead of F1) – 1st, 2nd and 3rd Specification

User behavior only

1st Specification

Best params:

{'penalty': 'l1', 'solver': 'liblinear'}
0.147633959638135
Predicting test outcome:
Classification report:

	precision	recall	f1-score	support
0	0.93	1.00	0.96	14368
1	0.75	0.16	0.27	1247
accuracy			0.93	15615
macro avg	0.84	0.58	0.62	15615
weighted avg	0.92	0.93	0.91	15615

Confusion matrix:
[[204 0 0 1043]
[0 0 0 0]
[0 0 0 0]
[69 0 0 14299]]
2nd Specification

Classification report:

	precision	recall	f1-score	support
0	0.94	0.97	0.96	17136
1	0.31	0.16	0.21	1247
accuracy			0.92	18383
macro avg	0.62	0.57	0.59	18383
weighted avg	0.90	0.92	0.91	18383

confusion matrix:
[[204 0 0 1043]
[315 0 0 1956]
[75 0 0 422]
[69 0 0 14299]]

3rd Specification

Best params:
{'penalty': 'l1', 'solver': 'liblinear'}
0.00834116214335421
Predicting test outcome:
Classification report:

	precision	recall	f1-score	support
0	0.93	1.00	0.96	17136
1	0.39	0.01	0.03	1247
accuracy			0.93	18383
macro avg	0.66	0.51	0.50	18383
weighted avg	0.90	0.93	0.90	18383

Confusion matrix:
[[18 0 0 1229]
[14 0 0 2257]
[8 0 0 489]
[6 0 0 14362]]

User behavior + LIWC

1st Specification

Best params:
{'penalty': 'l1', 'solver': 'liblinear'}

0.4232663535142658
Predicting test outcome:
Classification report:

	precision	recall	f1-score	support
0	0.95	0.99	0.97	14368
1	0.71	0.40	0.52	1247
accuracy			0.94	15615
macro avg	0.83	0.70	0.74	15615
weighted avg	0.93	0.94	0.93	15615

Confusion matrix:
[[504 0 0 743]
[0 0 0 0]
[0 0 0 0]
[201 0 0 14167]]

2nd Specification

Classification report:

	precision	recall	f1-score	support
0	0.96	0.93	0.94	17136
1	0.30	0.40	0.34	1247
accuracy			0.90	18383
macro avg	0.63	0.67	0.64	18383
weighted avg	0.91	0.90	0.90	18383

confusion matrix:
[[504 0 0 743]
[833 0 0 1438]
[153 0 0 344]
[201 0 0 14167]]

3rd Specification

Best params:
{'penalty': 'l1', 'solver': 'liblinear'}
0.04712421711899792
Predicting test outcome:
Classification report:

	precision	recall	f1-score	support
0	0.93	0.99	0.96	17136
1	0.39	0.05	0.09	1247
accuracy			0.93	18383
macro avg	0.66	0.52	0.52	18383
weighted avg	0.90	0.93	0.90	18383

Confusion matrix:
[[60 0 0 1187]
[77 0 0 2194]
[6 0 0 491]
[12 0 0 14356]]

LIWC only

1st Specification

Best params:
{'penalty': 'l1', 'solver': 'liblinear'}
0.3811560542797494
Predicting test outcome:
Classification report:

	precision	recall	f1-score	support
0	0.95	0.99	0.97	14368
1	0.71	0.38	0.49	1247
accuracy			0.94	15615
macro avg	0.83	0.68	0.73	15615
weighted avg	0.93	0.94	0.93	15615

Confusion matrix:
[[472 0 0 775]
[0 0 0 0]
[0 0 0 0]
[195 0 0 14173]]

2nd Specification

Classification report:

	precision	recall	f1-score	support
0	0.95	0.94	0.94	17136
1	0.30	0.38	0.34	1247
accuracy			0.90	18383
macro avg	0.63	0.66	0.64	18383
weighted avg	0.91	0.90	0.90	18383

confusion matrix:
[[472 0 0 775]
[757 0 0 1514]
[142 0 0 355]
[195 0 0 14173]]

3rd Specification

Best params:
{'penalty': 'l1', 'solver': 'liblinear'}
0.04420668058455114
Predicting test outcome:
Classification report:

	precision	recall	f1-score	support
0	0.93	0.99	0.96	17136
1	0.34	0.04	0.07	1247
accuracy			0.93	18383
macro avg	0.64	0.52	0.52	18383
weighted avg	0.89	0.93	0.90	18383

Confusion matrix:
[[48 0 0 1199]
[75 0 0 2196]
[8 0 0 489]
[10 0 0 14358]]