Janith Weerasinghe*, Kediel Morales, and Rachel Greenstadt

"Because... I was told... so much": Linguistic Indicators of Mental Health Status on Twitter

Abstract: Recent studies have shown that machine learning can identify individuals with mental illnesses by analyzing their social media posts. Topics and words related to mental health are some of the top predictors. These findings have implications for early detection of mental illnesses. However, they also raise numerous privacy concerns. To fully evaluate the implications for privacy, we analyze the performance of different machine learning models in the absence of tweets that talk about mental illnesses. Our results show that machine learning can be used to make predictions even if the users do not actively talk about their mental illness. To fully understand the implications of these findings, we analyze the features that make these predictions possible. We analyze bag-of-words, word clusters, part of speech n-gram features, and topic models to understand the machine learning model and to discover language patterns that differentiate individuals with mental illnesses from a control group. This analysis confirmed some of the known language patterns and uncovered several new patterns. We then discuss the possible applications of machine learning to identify mental illnesses, the feasibility of such applications, associated privacy implications, and analyze the feasibility of potential mitigations.

Keywords: Mental Illnesses, Machine Learning, Natural Language Processing, Privacy

DOI 10.2478/popets-2019-0063
Received 2019-02-28; revised 2019-06-15; accepted 2019-06-16.

1 Introduction

According to the World Health Organization, more than 300 million people suffer from depression [46] and an estimated 3.6% of adults in the United States had post-traumatic stress disorder (PTSD) in the year 2007 [19]. Better and more accessible diagnostic tools are needed because symptoms of mental illnesses are often missed and individuals may go undiagnosed.

Previous studies, both in psychology and linguistics, have shown that people with mental health disorders deviate from normal language use, and that these deviations can be used as a diagnostic tool. While early studies analyzed this relationship via patient essays and interview transcripts, recent studies have shown that similar changes in language usage can also be detected in social media posts. Moreover, more recent studies have shown that machine learning can predict the mental status of individuals through the content of their social media posts [17].

The 2015 ACL Workshop on Computational Linguistics and Clinical Psychology released a dataset containing tweets of users who had self-reported diagnoses of mental illness on Twitter [11]. While there are some limitations associated with the dataset we used, such as using self-reported diagnoses as the ground truth, which we discuss in Section 3.1, it still provides valuable insights and the release of this dataset has enabled the research community to build classifiers that detect mental illnesses with promising results. However, the ability of machine learning models to infer an individual's mental health status based on their social media posts raises numerous privacy concerns. In this work, we analyze several machine learning models that predict mental illnesses, identify the features that make these predictions possible, analyze privacy implications, and propose mitigations. Our main contributions are as follows:

– **Analyzing the impact of direct mentions of mental illnesses:** Given the self-reported nature of these datasets, some of these individuals may be actively talking about their mental illness on social media. To assess whether machine learning models pose a privacy threat, we must evaluate the performance of these models on instances in which users *do not* disclose their diagnoses. We measure classifier performance before and after tweets on mental health are removed from the data to investigate whether the models are simply detecting the active mentions of the mental illness or whether they are detecting more complex language patterns. Our re-

***Corresponding Author: Janith Weerasinghe:** Affil, E-mail: janith@nyu.edu
Kediel Morales: Affil, E-mail: km3556@drexel.edu
Rachel Greenstadt: Affil, E-mail: greenstadt@nyu.edu

BY-NC-ND

sults do not show a significant drop in prediction accuracy after statements of self-reported diagnoses are removed. Ours is the first study to show that it is possible to identify individuals with mental illnesses who do not explicitly discuss their diagnoses based on their social media posts.

– **Feature analysis:** From a privacy standpoint, it is important to understand what aspects of individuals' language use make them "classifiable." To identify predictive features of language usage, we conduct a feature analysis by applying feature importance measures to feature vectors created from word unigrams, part-of-speech tags, word clusters [29], and topic models [26]. We also use this feature analysis to aid us in designing evasion mechanisms.
  Our feature analysis confirmed some previously known language patterns (higher use of pronouns) exhibited by individuals with mental illnesses and we were able to identify several new language patterns (higher use of intensifiers, conjunctions and past participles).
– **Analysis of misclassifications:** We conduct a qualitative analysis to better understand instances in which the machine learning model misclassifies individuals. Understanding when and why the model fails is important for real-world deployment. We found that while some misclassifications are due to issues that could be rectified in future iterations, some false positives were due to more undesirable biases such as the users having similar background (military) or interests (music, sports) to those in the positive group. These words about military and war were identified as important predictors through our feature analysis as well.
– **Privacy implications, and mitigations:** We discuss the privacy implications of machine learning being applied to predict individuals' mental health status in several scenarios, and we discuss their feasibility. Through our analysis, we were able to show that machine learning can be used to identify, with a fair degree of accuracy, users with mental illnesses even if they are not actively talking about it, and identify linguistic patterns that distinguish individuals with mental illnesses. While it is possible to use a machine learning system as a first step in identifying mentally ill users, our work shows that due to undesirable biases that we uncovered in our misclassification analysis and feature analysis, and due to the low prior probability of mental illnesses in the general population and high false positive rates of the classifiers, similar classifiers should not be used as the sole input for detecting mental illnesses in the broader internet population. While this finding is reassuring from a privacy standpoint, misuses of this technology are still possible. Building upon this discussion, we suggest steps that can be taken by end users, social media platforms, and regulatory bodies to preserve the privacy of users, and we conduct experiments to demonstrate their feasibility.

2 Related Work

In Psychiatry and Psychology

Multiple studies in psychiatry, psychology, and medicine discuss linguistic differences between patients with mental illnesses and control groups. Mor et al. [28]'s meta-analysis showed that self-focus is associated with depression, anxiety, and negative mood. Zimmermann et al [47] showed that the relative frequency of first-person singular pronouns spoken in clinical interviews predicts the severity of future depressive symptoms. Smirnova et al. [40] found that Russian speakers with depression used more personal and indefinite pronouns and continuous and past-tense verbs. Van Der Zanden et al. [44] showed that "discrepancy words" (e.g: 'would', 'should', 'conflict', 'wish', and 'hope') used by depressed individuals predict treatment outcome and adherence. Al-Mosaiwi et al. [3] found that internet forums on anxiety, depression, and suicidal ideation contained more "absolutist words" (e.g: 'absolutely', 'all', 'always', 'must', 'never', and 'totally') than control forums.

In NLP and Machine Learning

Linguistic Inquiry and Word Count (LIWC) [42] is a widely used text analysis program that analyzes word counts based on a curated set of tokens. This program has been used to show differences in language use among students with neuroticism and depression [35], female assault victims being treated for chronic PTSD [21], and twitter users with PTSD [12]. Schwartz et al. [38] refer to this approach as a *closed-vocabulary* approach, as opposed to an *open-vocabulary* where words and features are discovered as part of the machine learning process.

There are multiple studies that have used social media data to predict depression and other mental illnesses. Guntuku et al. [17] comprehensively surveys these studies by comparing the data source, features used and the results. The study by De Choudhury et al. [13] was the first to use tweets to detect depression in

individuals. They used crowd-sourcing to identify Twitter users with depression via standard psychometric instrument (CES-D). They analyzed user engagement and emotion, different properties of social networks of the users, their linguistic style including depressive language use, mentions of antidepressants and symptoms, and words related to disclosure, relationships and life. They were able to achieve an accuracy of 69% with depression related language features and 68% with LIWC linguistic style features. They were able to achieve an accuracy of 72% when all the other features were included. These results illustrate that language is predictive of mental health status. Another foundational study by Coppersmith et al. [12] demonstrated that self-reported diagnoses of mental illnesses on social media can be used to create large datasets without needing to interview or survey subjects. They used a word unigram model, a character n-gram model, and a classifier that uses LIWC category frequencies to identify users with PTSD from a control group. Coppersmith et al. [9, 10] extended their earlier work by collecting and analyzing tweets of users having multiple mental illnesses including PTSD, depression, bipolar disorder and seasonal affective disorder (SAD). They showed that the classifiers based on the language models were much better at predicting mental illnesses than the classifier based on LIWC categories. This shows that there are words and language patterns that are not included in the LIWC categories that would help in identifying people with mental illnesses.

We use Coppersmith et al.'s dataset [11] containing tweets from users with depression or PTSD and tweets from a demographically matched control group. The aim of that work was to provide an apples-to-apples comparison of various approaches of modeling language relevant to mental health from social media. Several teams participated in this task. Resnik et al.'s model [34], which combined supervised topic modelling and bag-of-words features, had the best performance. Preoţiuc-Pietro et al. [32] used Differential Language Analysis (DLA) [38] to analyze language differences between individuals with mental illnesses and a control group. The work by De Choudhury et al. [13] also presented an analysis on the language use, social media engagement, and behavioral attributes that can be inferred from Twitter accounts of individuals with depression. Resnik et al. [33, 34] used supervised topic modeling to identify differences of content between depressed and non depressed individuals. From this set of studies emerged common sets of terms used more frequently by positive classes. In each study, these terms include those related to symptoms ('anxiety', 'withdrawal', 'severe', 'delusions'), treatment ('medication', 'side-effects', 'doctor', 'doses'), disclosure ('fun', 'play', 'helped', 'god') and relationships and life ('home', 'woman', 'she', 'him')[13]. Resnik et al. [34] in addition identified terms indicative of depression symptoms such as periods of low mood ('cry', 'crying') and low interest ('anymore', 'I used to'). Users suffering from PTSD too had used more terms related to treatment and terms like 'murdered', 'died', 'terrified' and 'anxiety'. All of these studies confirmed previous findings of heightened self focus of depressed individuals.

While some works[13, 32] explored only-positively correlated features with mental illness, negatively-correlated features too provided valuable insights. Resnik et al. [33, 34] showed topics on positive affect, social activities and family are negatively correlated with depression and topics on family and social activities are negatively correlated with PTSD.

At this point we would like to highlight how our work in language analysis differs from previous works on language analysis of individuals with mental illnesses. First, all of the previous works show that topics and words related to mental health are some of the top predictors. However, to fully evaluate the privacy implications, we need to examine the performance and the important features of these classification models in the *absence* of explicit mentions of mental health issues. Our analyses of tweets, after the removal of mental health-related content, closely models a scenario in which social media users do not explicitly reveal their mental health issues. In addition, we examine language whose usage is negatively-correlated with mental illness. We hypothesize that such features are important because a symptom of both depression and PTSD is the loss of interest in activities. Apart from De Choudhury et al. who analyzed the use of pronouns, studies have not focused on the syntactic differences in the language. Given that previous studies in psychiatry and psychology have found syntactical differences in language use, it is important to analyze and discover any syntactical differences in the language use. We aim to fill these gaps.

In Privacy and Ethics

Mikal et al. [27] conducted interviews with Twitter users to investigate user perspectives on the ethical issues surrounding the use of social media data for population-level depression monitoring. Their analysis revealed that Twitter users are generally aware that Twitter data is public by default and were not opposed to the use of publicly available data for health monitoring activities provided that the data are anonymized and aggregated

to remove personally identifiable information. Golder et al. [16] examined the attitudes of both social media users and researchers on using social media as a data source for research and had similar responses. Importantly, previous studies have also shown that users unwittingly share private information on Twitter, suggesting they may not be aware of what can be inferred from their publicly available data. For example, Mao et al. [25] showed that a classifier can identify tweets that leak private information such as vacation plans, medical conditions, and tweets made under the influence of alcohol. Sleeper et al. [39] and Wang et al. [45] show that both users on Facebook and Twitter at times regret the content that they have posted online.

However, we did not see any work that discusses in detail the privacy concerns of assessing a person's mental health status using their language on social media. Our work foremost contributes to the discussion of privacy and ethics in its detailed discussion of the privacy concerns of assessing a person's mental health status using their language on social media. In doing so, we build on Guntuku et al's [17] discussion of some such privacy concerns. Notably, Guntuku et al also highlighted the need for transparency about which health indicators are inferred by different parties involved and pointed out the open questions about the misclassifications in the previous studies, which must be understood before these models can be integrated into systems of care.

3 Experiments and Results

To determine the machine learning models' value as a diagnostic tool and their impact on privacy, we must determine the performance of the machine learning models in the absence of tweets that mention mental illnesses. Section 3.4 describes our approach to removing mental health-related tweets from the dataset then presents the performance of our machine learning models before and after mental health-related tweets are removed. We **did not observe** a significant drop in the classifier performance after mental health-related tweets were removed from the dataset. This suggests that there are other signals in one's language use that are predictive of mental health and may present a bigger privacy threat. Section 3.5 analyzes features to discover these other signals. Additional details to aid in reproducing this work are included in the appendix and the source-code of the experiments are available at https://github.com/janithnw/twitter_mh_public.

3.1 Dataset

We use a widely-used dataset from the 2015 ACL Workshop on Computational Linguistics and Clinical Psychology [11]. The dataset contains tweets from three types of users: users who have self-reported a diagnosis of depression, users who have self-reported a diagnosis of post-traumatic stress disorder (PTSD), and a demographically matched control group. A self-reported diagnosis is a tweet that contains a phrase similar to "I was diagnosed with depression" or "I was diagnosed with PTSD." These tweets were verified manually to remove jokes, quotes, or any other disingenuous tweets. For each verified tweet, the rest of the most recent 3000 tweets from the user who made the tweet (except the tweet with the diagnosis) were collected. As we will discuss in Section 3.4, the users may have tweeted about their mental illness more than once and such tweets may still be included in the dataset. The age and gender of each user was estimated, and a Twitter user with a similar age and gender was assigned as a matched control. Coppersmith et al. [12] contains more details about the procedures used in creating the dataset. The dataset contains 327 users with depression, 246 users with PTSD, and for each user with depression or PTSD, an age and gender-matched control user (1146 users in total). In our analysis, we are interested in two classification tasks: depression vs. control (**DvC**) and PTSD vs. control (**PvC**).

Although this dataset is powerful, it is also limited in multiple ways. The users captured in the positive class are users who self-reported their diagnosis, and therefore might have a different personality than users who are diagnosed with a mental illness and choose not to share their diagnosis on social media. A personality analysis conducted by Preoţiuc-Pietro et al. [32] on a similar dataset [9] shows that the language use of users in the positive class does overlap with the language predictive of personality traits such as openness, neuroticism, extraversion, and conscientiousness. However, it is hard to discern to what degree these personality traits are associated with their mental illness and their willingness to share their diagnoses on social media. Another limitation here is that depression and PTSD are complex diagnoses that manifest in a variety of ways and this dataset collapses and simplifies this reality. These issues are hard to avoid when using this approach to collect a large-scale dataset where it is not feasible to collect clinical standard ground truth.

3.2 Preprocessing

Before the tweets are analyzed, all re-tweets and tweets that contain URLs are removed because the content in these tweets was not written by the user. All mentions of usernames are replaced by "user". All Unicode emoji characters are converted to a text representation and all other Unicode characters are converted to ASCII characters. We use NLTK's TweetTokenizer [1] to tokenize the tweets. TweetTokenizer is a Twitter-aware tokenizer that treats ASCII emojis (such as ;) :-) <3) as a single token and limits the number of repeated characters to two (e.g. converts loool, looooooool to lool).

3.3 Features

We use four feature sets in our analyses. Most of the systems submitted to the CLPsych workshop [11] and other previous studies [17] show that bag-of-words features perform well. Resnik et al.'s system [34], which used a supervised topic modeling approach, performed best in the Shared Task. We therefore include these two feature sets in our analyses. Previous studies have shown that people with depression tend to use more personal pronouns [28, 47] and past tense verbs [40] in their writing. While De Choudhury et al. [13] included the frequencies of pronouns as a feature, differences in other part of speech constructs were not analyzed. Therefore we included part-of-speech (POS) tags in our analyses. One drawback of using sparse feature sets like bag-of-words is that the models could overfit and the analyses of these features become difficult. To overcome this we used clusters of related words. This creates a dense feature matrix and allows the model to generalize to previously unseen words.

Bag-of-Words: We use words that are used by more than 1% of the users as features. We did not remove function words as usually done in other natural language classification tasks because we wanted to detect potential differences between function word usage between the different populations.

Topic Models: We recreate the supervised topic modeling [26] approach that was used by Resnik et al. [34]. Following their approach, we build a 50-topic model by running LDA on stream-of-consciousness essays collected by Pennebaker and King [31]. These 50 topics were then used as informed priors for the Superviced LDA (sLDA) step. The sLDA model was trained on tweets from our dataset that were concatenated together based on the week that they were posted to form documents. Each user's label was used as the label for each weekly-aggregated document. To compute a single feature vector for each user, we compute the weighted average of the feature vectors for all the weeks, in which the weights are the fraction of tweets posted in a given week. We refer interested readers to Resnik et al. [34] for a detailed description of this approach. We did not conduct an extensive feature analysis for this feature set since Resnick et al. [34] and Preoţiuc-Pietro et al. [32] have performed a detailed feature analysis of LDA and sLDA topics on this dataset.

POS Tags: To discover grammatical-level differences in the language use of individuals with mental illnesses, we analyze uni, bi and tri-grams of part-of-speech tags. We use two part-of-speech Taggers (POS taggers). The first is a POS tagger that is trained on Twitter data from the Tweet NLP project [15] and is more accurate at tagging tweets. However, some of the POS tags are combined together to handle language usage patterns on Twitter and therefore are less descriptive than typical POS tags. For example the tag **L** is used for nominal proper noun and possessive verb combinations (examples: he's, I'm) and the tag **R** is used for all adverbs including comparative, superlative and wh-adverbs. The second, is the Python NLTK POS tagger [7] which uses the Penn Treebank tagset [37]. This tagger is meant for well-formed English sentences and therefore does not accurately identify POS tags in some tweets, but it distinguishes between types of pronouns and tenses of verbs. We noticed that emojis are not properly handled by POS taggers. Therefore we included an additional **EMJ** tag to represent emojis by replacing the POS tag given to emojis with **EMJ**.

Word Clusters: Word use on Twitter is informal and the same idea, word, or phrase can be expressed in different ways. For example, the tokens *I'll, Ima, imma* and *I'ma* mean the same thing and the words *quite, entirely, particularly, terribly* and *oddly* are semantically related. Clustering such words together and treating them as one token allows us to identify language patterns beyond simple word use and helps us discover more generalized language patterns. We use the set of 1000 hierarchical clusters created by Owoputi et al. [29] that are based on English tweets. They computed the clusters using Brown Clustering [8] which assigns words to classes based on the frequency of word co-occurrence resulting in a hierarchical set of classes that are grouped together semantically and syntactically. We replace words in tweets by their cluster identifier and remove words that do not belong to a cluster.

	Precision	Recall	F1
Depr. vs. Other	0.94	1.00	0.97
Depr. + Mental illnesses vs. Other	0.83	0.89	0.86
PTSD vs. Other	0.99	1.00	0.99
PTSD + Mental illness vs. Other	0.93	0.90	0.91

Table 1. 10-Fold cross-validated classifier performance for labelling direct mentions of mental illnesses in tweets

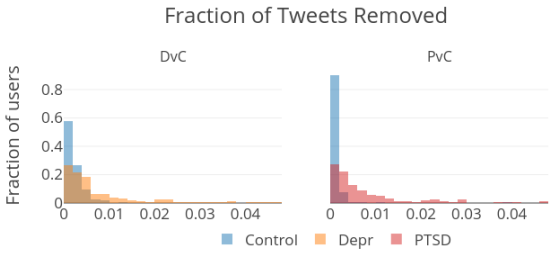


Fig. 1. Histograms of the fraction of mental illness related tweets removed by each user

3.4 Mentions of mental illnesses in tweets

Our initial hypothesis was that simple models such as bag-of-words use the active mentions of mental illnesses. To assess if machine learning models that detect mental illnesses pose a privacy threat, we must evaluate if these models can make successful predictions in the absence of active mentions of mental illnesses.

As mentioned earlier, the dataset that we use was collected based on self-reported diagnoses. If a user disclosed their diagnosis, all their tweets except for the diagnosis statement tweet are included in the dataset. Note that these users may have tweeted about their mental illness more than once and these tweets may still be included in the dataset. While reading through a random sample of tweets we realized that some users talk about their condition to raise awareness, to build a support network and to help other users with the same condition. In our dataset, 24% of users who have depression have mentioned the phrase "diagnosed with depression", and 33% from the PTSD set have the phrase "diagnosed with PTSD/P.T.S.D." or a similar phrase. None in the control group have tweeted a similar phrase.

We measure the effect such direct mentions of mental health have on prediction accuracy by measuring the accuracy before and after removing such tweets from the dataset. We identified three categories of tweets: direct mentions of depression (tweets about their diagnosis, or about depression in general) or PTSD, general mental health (tweets about other mental health illnesses, or mental well-being) and other tweets (all other tweets that do not belong to the previous categories).

An author labelled 3900 tweets from users with depression and 1000 tweets from users with PTSD (refer the appendix for a detailed labelling protocol). A second author labelled 500 tweets from each set to determine inter-annotator agreement (Cohen's kappa coefficient) which was 0.85 for depression labels and 0.96 for PTSD. We use machine learning to predict the labels for the other tweets. The tweets were preprocessed as described earlier and then tokenized by the NLTK TweetTokenizer. We use a Random Forest classifier with 500 trees on the following feature sets to make predictions:

– **Bag-of-Words:** TF-IDF values of words.
– **Word Clusters:** TF-IDF values of words belonging to a set of precomputed clusters. See section 3.3 for more details about this feature set.
– **Custom Word Lists:** We created four lists of words that are associated with depression, self-harm, suicide and PTSD such as *depression, mental-illness, P.T.S.D, suicidal, self-harm*. The presence or absence of words from each list in a given tweet was a feature. See appendix for the lists of words that were used.

Table 1 shows the performance of the classifier in detecting mental health-related tweets. To remove direct mentions of mental illnesses from the dataset we use the classifiers trained on the collapsed labels (i.e. depression + mental illness vs other and PTSD + mental illness vs other). To avoid introducing biases we filtered tweets of both the positive and the control users. Figure 1 shows the fraction of tweets removed from each user. On average 1.5% of tweets from users with depression and 7.9% of tweets from users with PTSD were removed. Users in the control group of the depression dataset had on average 0.3% of their tweets removed as well. Upon further inspection, most of these removed tweets were either talking about depression jokingly or used the words 'cut' and 'scars' which were used frequently in tweets about self-harm. Almost no tweets were removed from the control group of the PTSD dataset. In the proceeding sections, we will refer to the datasets with mental health-related tweets removed as *filtered* datasets.

Predictions Using the Filtered Dataset.
We use the different feature sets described in Section 3.3 on the two versions of the dataset (the filtered and the full version) to make predictions about the users' mental health conditions. Since we observed that a large

number of users did talked about their mental illness in their tweets, and that previous studies have shown mental health-related language was a top predictor for the positive class [13], our initial hypothesis was that we would observe a large drop in classifier performance in the filtered version. **However, we observed only a slight drop, as shown in Table 2**. These results suggest that after tweets that contain direct mentions of mental illnesses were removed, even simple classifier models are able to predict if a user has depression or PTSD. Our work is the first to show that simple classifiers can still predict mental health status after filtering out active discussion of mental illness.

For these experiments, we use a Support Vector Machine classifier with a linear kernel on various combinations of the feature sets. We did not change the default SVM parameters. While it would have been possible to achieve higher accuracy by tuning parameters, we opted not to perform any parameter tuning due to the limited size of our dataset which makes it difficult to test the model on a reasonably sized validation set.

Table 2 shows the results of the two predictions tasks—Depression vs Control (**DvC**) and PTSD vs. Control (**PvC**)—using 10-fold cross-validation. The relatively low standard error values for both AUC and average precision suggests that the performance of the models were stable and that effects of any overfitting of models are minimal [22].

Note that when performing cross-validation, the dataset is split on user-level (and not on tweet-level) and the predictions are made for an individual user. To construct the feature vector for a given user, their tweets are concatenated to compute the bag of words, word clusters and POS tag feature vectors. To compute the sLDA feature vector for each user, as described in Section 3.3, tweets are concatenated based on the week that they were posted and then weighted average of the sLDA output of each such weekly-aggregated document is taken as the feature vector for the user.

The best performing model for both classification tasks used the sLDA+BoW feature sets, which was also the best performing model in the CLPsych Shared Task workshop [11]. When considering most other feature combinations, adding the word clusters and part-of-speech tag feature sets improved the performance of the classifiers. For comparison, the systems submitted to the CLPsych Shared Task workshop had average precisions in the range of $0.74 - 0.87$ for the depression vs. control task and $0.72 - 0.89$ for the PTSD vs. control task. These systems used the complete dataset for training and were tested on a held-out dataset that we did not have access to. The survey by Guntuku et al. [17] state that the performance of primary care physicians in identifying depression, measured as AUC falls between $0.62 - 0.74$ and the performance on standard screening inventories are around 0.9. This suggests that machine learning systems have the potential of being used as diagnostic systems. We will discuss further in Section 4.1.

3.5 Feature Analysis

If our classifier is not relying on active mentions of mental illness to identify afflicted users, what other linguistic features distinguish these users? This section describes the analysis we conducted to determine which features the classifier is using to make these distinctions. This analysis will aid us in understanding the degree to which these classifiers pose a privacy threat, aid us in designing evasion mechanisms, and potentially illuminate novel aspects of how these illnesses affect language use.

We construct multiple feature sets and quantify the importance of each feature using information gain. We also measure the statistical significance and effect size of the term-frequency difference between the positive and the control classes for each feature. We report only this measure of feature importance, as we find the same trends when feature importance is measured through recursive feature elimination or as the mean decrease in accuracy as each feature is removed.

Analysis methods

We compute Term-Frequency (TF) and Term Frequency-Inverse Document Frequency (TF-IDF) values for each of the four feature sets. The TF values are used compute the statistical significance values and the Cohen's d measure and the TF-IDF values are used to compute the information gain.

Information Gain: Since the features are continuous, we compute the information gain for each feature as the maximum information gain obtainable by thresholding the feature at 100 points. Formally, given a feature X, the information gain of the feature when splitting it at a value t is given by $IG(X,t) = H(X) - H(X,t)$ where $H(X)$ is the entropy before the split and $H(X,t)$ is the entropy after the split. $H(X,t)$ is given by

$$H(X,t) = H\left(\frac{p^-}{p^-+n^-}, \frac{n^-}{p^-+n^-}\right) P(X < t) + H\left(\frac{p^+}{p^++n^+}, \frac{n^+}{p^++n^+}\right) P(X \geq t)$$

Model	AUC		AP		Precision		Recall	
	All	Flt	All	Flt	All	Flt	All	Flt
Depression vs. Control:								
BoW	0.821 ±0.012	0.810 ±0.013	0.830 ±0.016	0.821 ±0.016	0.76	0.74	0.72	0.71
BoW+Cl+TNLP	0.840 ±0.013	0.829 ±0.014	0.849 ±0.016	0.835 ±0.018	0.77	0.77	0.74	0.74
BoW+Cl+NLTK	0.836 ±0.015	0.826 ±0.016	0.841 ±0.021	0.830 ±0.022	0.78	0.76	0.75	0.74
sLDA+BoW	0.861 ±0.013	0.845 ±0.013	0.862 ±0.016	0.841 ±0.017	**0.81**	0.78	**0.76**	0.74
sLDA+BoW+Cl	0.862 ±0.016	0.846 ±0.015	0.865 ±0.018	0.850 ±0.018	**0.81**	**0.79**	0.75	**0.76**
sLDA+BoW+Cl+TNLP	0.858 ±0.018	0.841 ±0.018	0.858 ±0.023	0.842 ±0.023	0.80	0.78	0.74	0.73
sLDA+BoW+Cl+NLTK	**0.864 ±0.016**	**0.851 ±0.016**	**0.867 ±0.018**	**0.854 ±0.018**	**0.81**	0.78	0.75	0.75
PTSD vs. Control:								
BoW	0.838 ±0.015	0.824 ±0.018	0.854 ±0.015	0.839 ±0.018	0.79	0.76	0.73	0.72
BoW+Cl+TNLP	0.844 ±0.013	0.834 ±0.015	0.859 ±0.014	0.846 ±0.017	0.79	0.79	0.77	0.77
BoW+Cl+NLTK	0.840 ±0.013	0.829 ±0.015	0.856 ±0.014	0.840 ±0.016	0.78	0.78	0.77	0.77
sLDA+BoW	**0.868 ±0.009**	0.858 ±0.011	**0.883 ±0.009**	**0.873 ±0.011**	**0.81**	**0.81**	0.75	0.73
sLDA+BoW+Cl	0.863 ±0.010	0.855 ±0.010	0.874 ±0.011	**0.873 ±0.009**	0.79	0.77	0.78	0.76
sLDA+BoW+Cl+TNLP	0.861 ±0.011	0.851 ±0.011	0.875 ±0.012	0.867 ±0.012	0.80	0.79	**0.80**	**0.79**
sLDA+BoW+Cl+NLTK	**0.868 ±0.010**	**0.859 ±0.011**	0.876 ±0.013	0.871 ±0.011	0.80	0.78	0.79	0.77

Table 2. 10-fold cross-validated performance results for the two classification tasks, on the filtered (Flt) and non-filtered (All) datasets, when different combinations of bag-of-words (BoW), word clusters (Cl), TNLP POS tag (TNLP), NLTK POS tag (NLTK), and Supervised Topic Model (sLDA) features are used. The standard error for AUC and Average Precision (AP) are also shown.

Here, p^- and p^+ are the number of positive samples when $X < t$ and $X \geq t$, respectively, and n^- and n^+ are the number of negative samples when $X < t$ and $X \geq t$. The information gain of the feature is the maximum value that can be obtained for $IG(X,t)$ at 100 t values within three standard deviations of the mean.

Statistical approaches: We conduct t-tests for each feature to determine if there is a statistically significant difference between positive and control populations. Because we are comparing multiple features at once, we apply Bonferroni Correction [14] to avoid the problem of multiple comparisons as done by Choudhury et al. [13] and Schwartz et al [38]. The Bonferroni Correction is performed by dividing the significance threshold by the number of features compared. After this correction, the significance threshold of 0.05/23000 for Bag-of-Words, 0.05/8000 for TweetNLP, 0.05/22000 for NLTK, and 0.05/1000 for Word Clusters. As pointed out by many studies [23, 41], using statistical significance alone when the number of samples is large is not informative because even small differences between the populations tend to become significant. We therefore report Cohen's d as a measure of the size of the difference between the two populations. The Cohen's d value was computed as: $d = \frac{\mu_p - \mu_n}{s}$, where μ_p is the mean of the frequency of the feature in the positive class, μ_n is the same value for the negative (control) class, and s is the pooled standard deviation. Generally, if the absolute value of Cohen's d is between 0.2 and 0.4, it is usually considered to be a small effect, a value between 0.4 and 0.8 is considered a medium effect and a value greater than 0.8 is considered to be a large effect.

One limitation of this analysis is that Cohen's d measure assumes that the variables are normally distributed. Most of the features in our analyses follow a skewed distribution because these features have zero term frequencies for a large portion of users. This assumption is of less concern in our context, however, because Cohen's d measure *underestimates* the effect size for skewed distributions, and furthermore, this underestimation is minimal for large sample sizes [20, 36]. The t-test assumes that the test statistic—in our case the sample mean—is normally distributed. According to the central limit theorem, for large sample sizes, the sample mean is normally distributed even if the underlying distribution is not normal. Because we are using a large sample size, the effects due to the skewness of the distribution should therefore be negligible.

Bag-of-words features

Table 3 shows the top 20 bag-of-words features with the highest information gain for the two classification tasks, with and without mental health-related tweet filtering. Figure 2 shows all the significant features with an effect size more than 0.2 represented as a word cloud. Some of the highly-ranked features may be artifacts of the time at which the data was collected such as "ebola"

(The dataset was collected during peak of the Ebola outbreak) and "5sos" (5 Seconds of Summer, an Australian pop rock band). On the other hand, the statistically significant features distinguish the different classes across the whole population. For example tokens such as "I've" and "myself" are more general features that distinguish between the two classes. Figure 2 shows that individuals with depression tweeted fewer abbreviations such as "lol" and "smh", and tweeted less about general day-to-day things (less use of "ready, early, Monday, nap"), and show increased use of swear words and conjunctive words such as "and" and "apparently." These patterns become more apparent in feature analysis of word clusters and POS tags. Similarly individuals with PTSD used a higher number personal pronouns ("I've") and words like "being", "also", and "through."

Task	Top 20 features
DvC, All Tweets	diagnosed, depression, anxiety, ebola, basketball, meds, suicidal, **ready**, #, **lol**, **and**, severe, 5sos, **fucking**, harm, pandora, emj:thumbsdown, fandom, <, **!**
DvC, Filtered Tweets	ebola, basketball, come, #, **ready**, fandom, **and**, **lol**, meds, **fucking**, 5sos, pandora, **!**, wale, emj:angryface, **ya**, #missyou, weekend, miss, fucking
PvC, All Tweets	ptsd, **diagnosed**, depression, diagnosis, ebola, suicide, meds, #ptsd, **being**, **anxiety**, disorder, harm, **also**, isis, trauma, **mental**, **weekend**, assault, **therapist**, illness
PvC, Filtered Tweets	ebola, **being**, meds, harm, isis, **also**, ready, caring, assault, because, racism, weekend, military, awful, therapist, trauma, assholes, warrior, gender, abuse

Table 3. Top 20 bag-of-words features with the highest information gain. Words in **boldface** are significant at Bonferroni corrected $p < 0.05$ and Cohen's $d \geq 0.2$

(a) Depression vs. Control **(b)** PTSD vs. Control

Fig. 2. Statistically signficant Bag-of-Words features. Size represents information gain. Red words indicate the positive class and blue words indicate the control class.

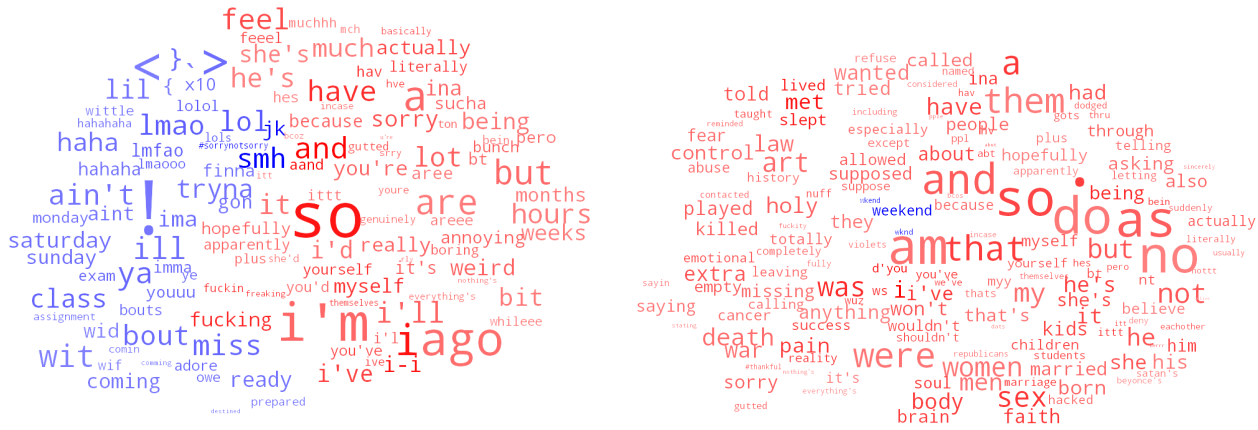
Word clusters

Figure 3 shows the significant word clusters from the *filtered* datasets for the two classification tasks as a word cloud. Users in the depressed class used fewer internet abbreviations and hash-tags associated with positive sentiments (*lol, lmao, #sorrynotsorry*), word clusters that describe day-to-day activities, and family and friends. They also used a higher frequency of words describing personal attention (*I, myself*), second person pronouns (*you're*), and swear words. While increased usage of personal attention-related words and swear words have been linked to depression by previous studies, clusters with words such as "and" and "because" have not been identified as markers of depression by previous studies. We also find that clusters of intensifier words such as "lot", "ton", "so" and "much" and clusters that include words such as "ago", "have" and "actually" are used more frequently by depressed individuals.

Similarly, in the PTSD-control feature set, we see reduced usage of daily activity-related words, increased use of conjunctions such as "and" and "because", clusters about personal focus, and increased use of personal and second person pronouns (*I've, you've, myself*) among users with PTSD. Word clusters on topics such as sex, faith and marriage, kids, law and history also occurred more frequently in tweets by individuals with PTSD. Clusters about war, death, kill and suicide may be predictive of PTSD because the condition is common among those with a military background.

Part of Speech Tags

The TweetNLP tagger correctly tagged Twitter specific language but its POS tags were not as specific as the NLTK POS tagger. The results from the two taggers thus provide complementary insights. Table 4 shows the statistically significant POS tag n-grams that had the highest effect sizes. The usage examples are the most frequently occurring phrases in the text for each tag n-gram. For the DvC classification task, the TweetNLP POS tags produced more significant tag n-grams. N-grams with nominal+verbal(**L**), adjective(**A**) and adverb(**R**) tags, conjunctions (**&**) together with common nouns(**N**) and pronouns (personal and wh–)(**O**) had higher information gain and high effect sizes. This analysis also reveal that Twitter-specific discourse markers such as ":" and hashtags were used significantly less by the depressed users. Analysis using the NLTK POS tags showed higher use of conjunctions (**CC**) alongside singular(**NN**) and plural nouns(**NNS**) in users with depression, and increased conjunction use also occurred



(a) Depression vs. Control **(b)** PTSD vs. Control

Fig. 3. Word cloud of significant clusters (Bonferroni corrected $p < 0.05$ and absolute value of Cohen's $d \geq 0.2$). For each cluster the top 10 tokens are grouped together. The size represents the information gain and the color represents the valence and Cohen's d value - red shades show a higher positive Cohen's d value indicating that the cluster is more frequent in the positive class and a blue shade is associated with the control group.

among users with PTSD. Users with PTSD used more personal pronouns (**PRP**), past tense (**VBD**), and past participle verbs (**VBN**) than control users.

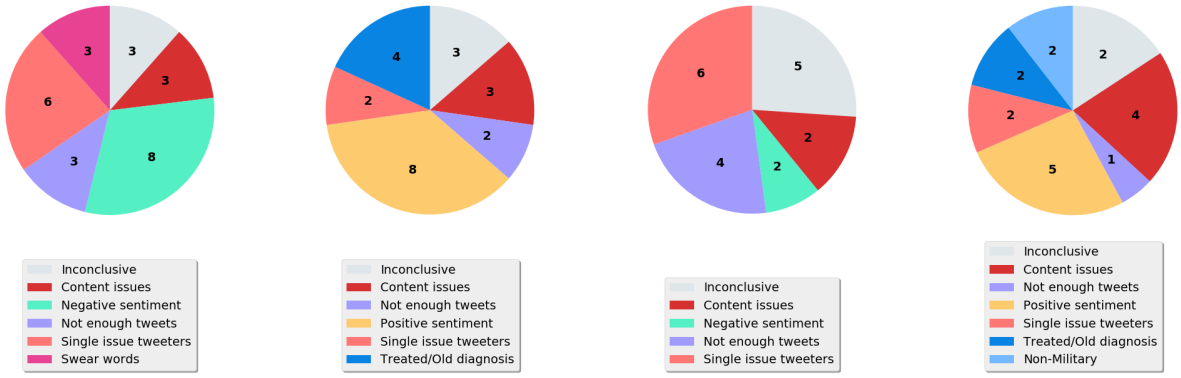
Topic Models

Table 5 shows the five topics with the greatest information gain for the filtered datasets. In accordance with results from the word clusters, these topics include positive sentiment, swear words, and day-to-day activities. We did not observe a significant difference between the topics learned from the filtered and unfiltered datasets. We believe that since the priors for the topics were computed using the stream-of-consciousness essays, any new topics such as topics relevant to mental health disclosure were not learned by these models.

3.6 Misclassification Analysis

In the filtered dataset, both the false positive and false negative rates of the models for DvC ranged $0.23 - 0.28$, and for PvC, the false positive rate was $0.28 - 0.29$ and the false negative rate was $0.23 - 0.27$. To understand the biases of the machine learning model in detail and identify the reasons for misclassifications, we built a classifier that uses the bag-of-words feature set. After running this classifier on the filtered data, we manually analyzed the misclassifications using three techniques. We created a word cloud from each user's tweets to get an overview of the user's Twitter feed. For false negative instances, we searched for depression/PTSD related words in the unfiltered tweets of each user to observe what they mention about their disease. We computed the impact of each individual tweet toward the prediction probability. Informally, the impact of each tweet quantifies how much an individual tweet contributes towards the classifier decision. Formally, given a previously trained model M, a user's set of tweets T, and a function prob that outputs the classifier probability given a model and an instance, we computed the impact of each tweet $\texttt{impact}(M, t_i)$ for $t_i \in T$ as $\texttt{prob}(M, T) - \texttt{prob}(M, T - \{t_i\})$. We inspected the top 10 tweets by impact—tweets that when removed, caused the prediction probability of the positive class to drop the most—and the 10 tweets that had the lowest (or the highest negative) impact—tweets that when removed, caused the prediction probability to increase the most. We will refer to the former as "positive-leaning tweets" and the latter as "negative-leaning tweets." The positive leaning tweets provided valuable insights. Most negative leaning tweets were tweets with emojis and positive social media abbreviations like 'lol'.

Figure 4 shows the most likely reasons for the misclassifications. First, several had very few (< 70) valid tweets (valid tweets are tweets that do not contain URLs or retweets). Several other users had content that differs from the rest of our dataset such as tweets with a large number of Emoji characters, some non-English tweets and repeated content most likely generated by other



(a) **DvC** False Positives (b) **DvC** False Negatives (c) **PvC** False Positives (d) **PvC** False Negatives

Fig. 4. Most likely reasons for misclassifications for the two classification tasks. The numbers on the pie chart represent the number of misclassified instances in each category.

Tags	IG	C*d*	Usage
DvC - TweetNLP Tagger:			
&	0.050	0.45	and, but
L R	0.043	0.46	I'm not, I'm so
L R A	0.044	0.46	I'm pretty sure, I'm so excited
A N &	0.048	0.41	last night and, best friend and
& L	0.036	0.41	and I'm, but I'm
@ L	0.032	0.37	@user I'm not, @user I'm so
L R R	0.036	0.39	I'm not even, it's not even
∼ #	0.058	-0.43	: #mentionto, : #inmiddleschool
∼ V V	0.042	-0.45	: can't wait, : don't get
DvC - NLTK Tagger:			
NN CC	0.046	0.40	friends and, home and
CC	0.042	0.41	and, but
NNS CC	0.057	0.38	@user but, people and
PvC - TweetNLP Tagger:			
& R	0.045	0.47	and then, but not
A &	0.028	0.48	strong and, sleepy and
O V A	0.044	0.47	you are right, I am sorry
&	0.037	0.45	and, or
PvC - NLTK Tagger:			
PRP VBD VBN	0.075	0.58	I was told, I was gonna
PRP VBD	0.038	0.60	I was, it was
CC RB	0.045	0.58	and then, but not
VBD VBN	0.053	0.49	had been, was told
CC	0.051	0.54	and, but
PRP VBD TO	0.042	0.49	I had to, I used to
PRP	0.033	0.50	I, it
VBD	0.038	0.49	was, got
IN PRP VBD	0.058	0.47	if I had, that I was

Table 4. Selected POS Tag N-grams with high information gain and effect size – Cohen's *d* (**C*d***) that are statistically significant for PvC and DvC classifications

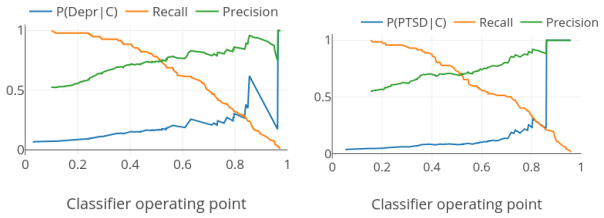
IG	Topic Words
DvC:	
0.012	ur, bc, da, dat, #, money, dey, ppl, wat, lmao
0.005	dont, cant, ill, didnt, ive, aint, doesnt, wait
0.003	tonight, tomorrow, week, weekend, night, wait
0.002	love, you're, girl, best, baby, girls, beautiful
0.002	bitch, fuck, ass, shit, lmao, bitches, fuckin
PvC:	
0.026	da, ain't, dat, tho, ass, cuz, shit, bitch
0.025	coffee, honestly, black, face, daily, white
0.020	day, today, happy, national, birthday, holiday
0.015	win, team, year, season, good, tonight, fans
0.010	haha, hahaha, yeah, bout, sooo, wtf, good, bad

Table 5. Statistically significant (Bonferroni corrected $p < 0.05$) topic models with the higest information gain values

applications (for example: tweets mentioning daily follower/unfollower counts, tweets announcing new Facebook photo uploads). These instances were labelled as having "Content issues." We believe some of the false positives in the **DvC** classification task were caused by the users posting a large number of swear words. Several other users were classified as having depression or PTSD due to tweets that are more self-focused, angry tweets, and tweets about relationships and the hardships of life. All such instances were grouped as 'Negative sentiment' in Figure 4. There were several misclassified users who primarily talked about a single topic such as politics, sports, musicians and bands, religion and health issues. Such instances were labelled as "Single issue tweeters." Interestingly some of the false negatives for both depression and PTSD classes included individuals who seem to be either on medication (iden-

tified by observing tweets mentioning therapy or antidepressants), who were already treated (identified by tweets mentioning their mental illness in the past tense), or who reported an older diagnosis. Other false-negative users had tweeted positive sentiment words such as 'lol', 'lmao' and 'love', which may have resulted them being classified as negative. Two users with PTSD but without a military background were classified as negative suggesting that the PTSD classifier is relying heavily on the correlation between military service and PTSD, rather than other linguistic patterns. One "single issue tweeter" with a large number of military-related tweets was in the control group, but was classified as having PTSD. We were unable to identify reasons for the misclassifications marked as "Inconclusive" in Figure 4.

Some of these misclassifications, especially the ones due to content issues, could be avoided by modifying the prepossessing step. However, we did not incorporate such modifications to our work since that could result in a preprocessing stage that is overfit to this dataset.



(a) Depression vs. Control **(b)** PTSD vs. Control

Fig. 5. Posterior probability, precision, and recall (y-axes) values when changing the probability threshold (x-axes) of the classifier.

4 Applications, their Feasibility and Privacy Implications

For many individuals suffering from mental illnesses, social media is a safe space to express themselves, network, and encourage one another. In this section, we analyze the feasibility of using machine learning to identify individuals with mental illnesses and how different parties can mitigate privacy-invasive actions.

4.1 As a Diagnostic Tool

Our results show that even after removing direct mentions of mental illnesses, simple machine learning algorithms are able to predict users suffering from a mental illness with a fair degree of accuracy. As a result, users who have not revealed their mental health diagnosis or users who have not been diagnosed could be identified through the analysis of their social media posts. As mentioned earlier, the performance of current classifiers matches or exceeds the performance of primary care physicians to detect depression. However, standard screening interviews designed to diagnose depression perform better than the classifiers [17]. Therefore, there is a real possibility that social media screening can serve as a first step in identifying an individual's mental illness and directing them towards professional help. Social media platforms or other responsible parties can provide proactive help to users who potentially have mental illnesses. However, even such well-meaning endeavours need to be carefully implemented so that they do not invade individual privacy, and any interventions are done in a sensitive manner. The best example for a scenario where a well meaning application faced backlash from the community due to privacy concerns is the Samaritans Radar Twitter app [24]. It monitored a user's Twitter contacts for phrases like "help me" and "hate myself" and flagged individuals as are struggling to cope or suicidal and then offered suggestions of how to approach them and offer help. Twitter is used by many individuals to form support networks where they can open up about their struggles. Many users criticized the app for invading people's safe spaces and potentially allowing trolls and bullies to target individuals when they are most vulnerable.

Our analysis of misclassifications revealed some of the undesirable instances of false-positives such as individuals being identified as having depression because they tweeted about music and bands that are mostly talked about by depressed individuals, and individuals who tweeted about military being classified as having PTSD. Before a diagnostic tool is deployed, steps needs to be taken to avoid these types of misclassifications.

4.2 Unethical and Malicious Uses

Although it is understood that tweets posted on Twitter are publicly visible by default, most individuals are unaware of the associated privacy leaks[25] and at times regret things they've posted[39, 45]. The possibility of using social media text to assess one's mental health status raises privacy concerns, especially given the revelations that people's psychological profiles were used to target specific advertisements to them during the 2016 US presidential election. Since social media platforms

such as Twitter and Facebook allow advertisers to create target audiences by specifying a list of user identifiers[1], it is possible for advertisers with malicious or unethical intentions to create a list of target users who may have depression or PTSD. While custom audiences are a valuable tool for advertisers to reach their target customers and have a lot of legitimate uses, Andreou et al. [5] have shown that Facebook users receive ads that are targeted using invasive strategies. Currently, there are no effective technical solutions to prevent the misuse of this feature. While both Twitter[2] and Facebook[3] prohibit ad targeting of this nature, it is unclear how these policies can be enforced.

Data brokers also have an interest in the mental health data of individuals. These companies aggregate and analyze personal information about consumers from a variety of sources and share them with other parties for purposes such as marketing products, verifying an individual's identity, or detecting fraud. There is little transparency and accountability on what type of data are gathered by data brokers and how these data are shared [43]. Reports show that data brokers collect information from public sources including social media sites and use it to make inferences about multiple fields including health data [43]. Employers and insurance companies are other parties who have potential interest in such data.

4.3 Feasibility

While social media posts can be informative about the mental health of an individual, they should not be used as the sole input to definitively determine if a user is suffering from a mental illness. The current predictions are performed on datasets that have a higher number of positive instances than the general population. Given the false positive rates of current systems it is likely that there will be a large number of false positives. We use Bayes Theorem and the prior probability of the prevalence of a given mental illness in the general population to estimate the probability of an an individual actually having the mental illness given that the system classified them as positive. Let Dpr be the event that an individual actually having depression and let C be the event that the classifier predicts that the individual has depression. Then, using Bayes Theorem, the probability that an individual has depression given our classifier predicted as such (Bayesian Detection Rate) is given by:

$$P(Dpr|C) = \frac{P(C|Dpr)P(Dpr)}{P(C|Dpr)P(Dpr) + P(C|\neg Dpr)P(\neg Dpr)}$$

$P(C|Dpr)$ is given by the True Positive Rate (Recall) of the classifier and $P(C|\neg Dpr)$ is given by the False Positive Rate. In 2016, 6.7% of all U.S. adults reported at least one major depressive episode [2], and this value can be used as the prior probability $P(Dpr)$. By using the performance values for our $sLDA+BoW$ model$P(Dpr|C) = \frac{0.781\times0.067}{0.781\times0.067+0.287\times(1-0.067)} = 0.163$ Similarly, the probability for PTSD, $P(PTSD|C)$, can be computed using 3.6% [19] as the prior probability, which results in a posterior probability of 0.09. These probability values are very low to be practically used to create a list of individuals with mental illnesses. The reason for such a low number is the low prior probability of mental illnesses in the general population and the high false positive rates. However, the number of false positives can be reduced by only selecting users that were predicted as positive with a high classifier confidence. Figure 5 shows how the posterior probability, precision, and recall changes when changing the probability threshold of the classifier. These graphs show that the posterior probability can be raised to around 50% by only considering predictions with a high confidence. However, this comes at a significant loss of recall. However, our misclassification analysis showed that multiple users who were misclassified into the positive class exhibited similar behaviours as those with mental illnesses, such as expressing anger and negative mood. Though these users may not be suffering from a mental illness, users with a similar mental state might still be of interest to advertisers with unethical intentions.

This phenomenon of getting a low posterior probability for an event is common in other domains such as intrusion detection systems [6] where the prior probability of an event is very low. This may seem contradictory to our earlier suggestion that the classifier performance matches that of primary care physicians. However, in clinical settings a primary care physician's initial diagnoses would be followed up by secondary screenings. Predictions made by the machine learning systems should not be used as the only input in making a diag-

1 Facebook custom audiences: https://www.facebook.com/business/products/ads/ad-targeting
Twitter tailored audiences: https://business.twitter.com/en/targeting/tailored-audiences.html

2 https://business.twitter.com/en/help/ads-policies/other-policy-requirements/policies-for-conversion-tracking-and-tailored-audiences.html

3 https://www.facebook.com/policies/ads/prohibited_content

nosis, instead they could be used as the first step or one source of input in a diagnostic system.

4.4 Mitigations

We showed that it is hard to use machine learning to accurately predict the mental health status of people from the general population due to the low prior probability of mental illnesses in the general population and the high false positive rate of the classifiers. However, this does not mean abuses of these techniques are not possible. In this section, we will discuss several mitigatory steps that different stakeholders involved can take to identify and/or prevent the usage of machine learning to target individuals with mental illnesses.

DvC Positive Leaning Tweets
USER Ok I'm crying so much right now :') f**k f**k f**k f**k f**k f**k f**k
PvC Positive Leaning Tweets
I'm not a f***ing role model. I'm a 21 year old emotional coaster with pipe dreams I don't understand why people brag about how much they can drink!!!
DvC Negative Leaning Tweets
USER 😘😘😘😘😘😘😘 USER you actually can if you want hahaha 😃😃😃😃😃😃
PvC Negative Leaning Tweets
Good morning and Good day!!! USER 🙈😂😂😂

Table 6. Examples of positive and negative tweets. These tweets were modified to protect user privacy.

End User Mitigations.
The results from our feature analysis and the misclassification analysis serve as a guide to how classifier predictions can be changed. For example, both analyses show that having a higher number of social media abbreviations and content with positive sentiments may push the classifier towards a negative prediction. They also show that simple models like the bag-of-words model tend to be brittle. Based on these analyses, we hypothesized that we could alter classifier predictions by adding, removing, or replacing a small handful of tweets.

We select the sLDA + BoW and BoW + Clusters + TweetNLP models to be used in this analysis because they cover the different feature sets that performed well in the prediction task. We train them on 70% of the filtered dataset, and to evaluate whether a positive prediction given to a user can be flipped by adding, removing, or replacing tweets, we select a random sample of users who had more than 1500 tweets. To remove tweets, for each selected user, we compute the impact that of each of their tweets had towards the prediction probability as discussed in Section 3.6. Then we remove tweets one-by-one, starting with the most positive leaning tweets and ran the prediction step again until the prediction changed. To determine the feasibility of adding tweets to flip a prediction, similar to the previous experiment, we add negative leaning tweets to each user until the prediction changed. For this experiment, the negative-leaning tweets were selected as follows: for each control user in the test set that had more than 1500 tweets, we compute the impact of each of their tweets towards their prediction probability and order all the tweets with the most negative learning tweets ordered first. These tweets primarily consisted of emoticons and abbreviations signifying positive sentiment. We then add each of these tweets one-by-one and measure the number of tweets that need to be added to change the prediction of a user. We also test a combination of these two approaches by replacing each of the removed positive-leaning tweets with a negative-leaning tweet. Figure 6 shows the number of tweets as a fraction of their initial tweet count that had to be added, removed, or replaced from users to flip their prediction, plotted against the classifier's initial probability.

Figure 6 shows the results of these experiments. Adding negative leaning tweets was more effective than removing positive leaning tweets and replacing tweets is the most effective approach. Predictably, it is harder to change the prediction for users that had a high classifier confidence. In most cases, the sLDA+BoW model was more robust against the addition or removal of tweets. Two possible reasons for this are that the topics distributions may not change by an adequate amount after removing individual tweets and the weak aggregation is robust against removing tweets. **However, the tweet replacement strategy works well on both models and the initial probability given to the user had a smaller impact on this strategy.** The average fraction of tweets that had to be replaced to flip the prediction was $3\% - 4\%$ for users with depression and $6\% - 7\%$ for users with PTSD.

For this sort of a mitigation strategy to be applied, a user requires access to a trained model. Here we assumed that the user has access to a trained model, but not one that included their own data. Such a trained model could be made available publicly or could be

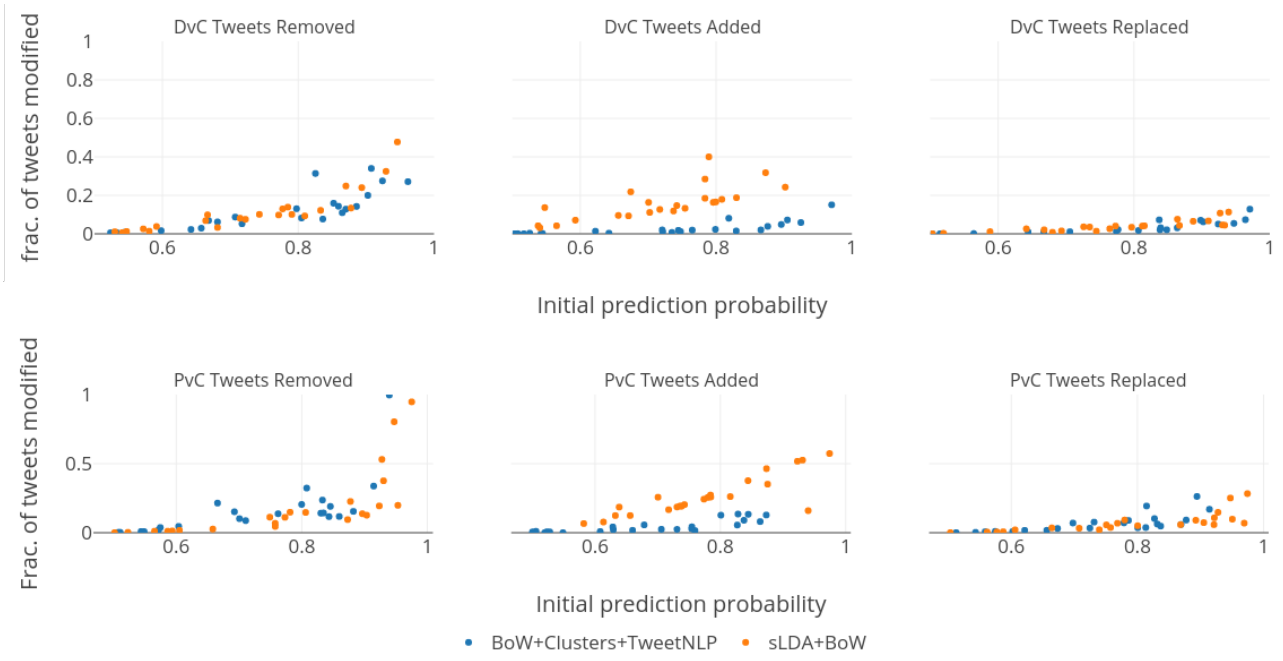


Fig. 6. The fraction of tweets that had to be added, removed or replaced to flip a positive prediction.

made into a service that shows a user's positive and negative leaning tweets. One limiting assumption we made here is that an adversary would be using the same model to make predictions. There is some evidence that due to the *transferability* property in machine learning models, adversarial samples produced to mislead one model can mislead other models [30]. Another mitigation strategy could be to flag or remove tweets based on a set of heuristics designed based on the insights we gathered from observing positive and negative leaning tweets, our misclassification analysis, and our feature analysis. Examples for such heuristics would be to flag tweets with swear words, tweets with high negative or angry sentiment. Similarly, our previous results can be used to create a "negative-leaning tweet generator" that would generate tweets containing social media abbreviations and emojis that would push a classifier towards a negative prediction. Creating such a heuristics based mitigation method, and validating it on a separate dataset would be an interesting avenue for future work.

This approach of altering the inputs to a machine learning or optimizing system to protect individuals or communities is an example of Protective Optimization Technologies (POTs) introduced by Gurses et al. [18]. POTs analyze different events that affect users and manipulate them (for example, by poisoning system inputs) to influence system outcomes in a favorable manner. Ideally, users of a social media platform should be able to express themselves freely and without the fear of being targeted because of their mental health status. Seeking out support on social media has been shown to be helpful to people dealing with depression [4] and we do not want this research to have the effect of silencing users or chilling their expression. We explored both additive and subtractive mitigations so that users concerned about keeping their mental health status private have multiple options. That said, there is also evidence to suggest that sometimes users regret social media posts made in emotionally "hot" states [45], and being able to warn users about such tweets might be beneficial.

Mitigations by Platforms and Regulatory Bodies. While social media platforms have policies in place that prohibit advertisers from targeting users based on their mental health, they do not use technical solutions to track such behavior or enforce these policies. Platforms could use a classifier that can predict mental illnesses to detect if a target audience list created by an advertiser consists of a significantly large number of users that are predicted as being positive by the classifier. However, such an approach, which evaluates the mental health of social media users, could be problematic unless it is deployed with safeguards to ensure that the users' mental health status is not exposed and user privacy is guaranteed. Having platforms run these analytics on their users might be as or more invasive than having advertisers do so, especially with the low degree of public trust in these

platforms. If this were to be done, structures would need to be put in place to insulate the employees and systems responsible from other business units and ensure that data from the processes is not retained. Similar approaches might be useful to regulators investigating suspected abuses by the platforms or third parties.

5 Discussion and Conclusions

Our findings show that individuals with depression or PTSD can be identified by analyzing their tweets even if they do not explicitly talk about their mental illness. Given this finding and its privacy implications, it is important to understand which factors of one's language use makes one "classifiable" as having a mental illness. To answer this question, we analyzed the important features and misclassified instances.

Our feature analysis corroborated some depression-linked language patterns identified previously such as higher levels of self-focus among users with depression [28, 47]. We saw similar results in users with PTSD. We were also able to discover several language patterns that were not identified in previous studies. Users with depression and PTSD used coordinating conjunctions such as *and*, *but*, and *or* more frequently than the control group. Users with depression also used more intensifiers, and fewer abbreviations associated with positive sentiment. Users with depression and users with PTSD both tweeted less about day-to-day activities.

Our qualitative analysis of misclassifications revealed insights about the classification model. Some false positives were due to language use exhibiting more self-focus, anger, and frustration. Other false positives related to interests that were shared mostly by the positive class (such as music bands, artists, and the military). These false positives demonstrate the limitations of deploying similar machine learning systems in the real world. Interestingly, some of the false negatives were from people who have had depression in the past but had likely since recovered. The analysis of the PTSD vs control classification task seems to show that the classifier associated military-related content with PTSD. To validate this hypothesis and to avoid such biases, individuals in the control class need to be matched more closely to those in the positive class.

Our analysis of applications and privacy implications showed that, while automated classifiers could be used as a first step in detecting a mental illness, they are not accurate enough to be used as the sole factor in determining the mental health status of an individual. This is reassuring, because these accuracy limitations make it more difficult for someone to automatically target individuals with mental illness on social media platforms. However, despite the potential inaccuracies of these classifications, we point out many possible ways in which these tools could be abused. We therefore also suggest and validate some potential defenses which add, delete, or replace tweets to alter classifier predictions.

Acknowledgements

We would like to thank the anonymous reviewers for their helpful comments and feedback, Cynthia Gill and Steven Rood-Ojalvo for their contributions in the early stages of this project, Glen Coppersmith for providing us the CLPsych dataset, and James Pennebaker for providing us the stream-of-consciousness essays dataset. Our work was supported by the National Science Foundation under grants 1253418 and 1931005.

References

[1] http://www.nltk.org/api/nltk.tokenize.html.

[2] https://www.nimh.nih.gov/health/statistics/major-depression.shtml.

[3] Mohammed Al-Mosaiwi and Tom Johnstone. In an absolute state: elevated use of absolutist words is a marker specific to anxiety, depression and suicidal ideation. *Clinical Psychological Science*, January 2018.

[4] Nazanin Andalibi, Pinar Ozturk, and Andrea Forte. Sensitive Self-disclosures, Responses, and Social Support on Instagram: The Case of #Depression. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, pages 1485–1500, New York, NY, USA, 2017. ACM.

[5] Athanasios Andreou, Marcio Silva, Fabrício Benevenuto, Oana Goga, Patrick Loiseau, and Alan Mislove. Measuring the Facebook Advertising Ecosystem. In *NDSS 2019 - Proceedings of the Network and Distributed System Security Symposium*, San Diego, United States, February 2019.

[6] Stefan Axelsson. The Base-rate Fallacy and Its Implications for the Difficulty of Intrusion Detection. In *Proceedings of the 6th ACM Conference on Computer and Communications Security*, CCS '99, pages 1–7, New York, NY, USA, 1999. ACM.

[7] Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python*. O'Reilly Media, Inc., 1st edition, 2009.

[8] Peter F. Brown, Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. Class-based N-gram Models of Natural Language. *Comput. Linguist.*, 18(4):467–

479, December 1992.

[9] Glen Coppersmith, Mark Dredze, and Craig Harman. Quantifying Mental Health Signals in Twitter. *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 51–60, 2014.

[10] Glen Coppersmith, Mark Dredze, Craig Harman, and Kristy Hollingshead. From ADHD to SAD: Analyzing the Language of Mental Health on Twitter through Self-Reported Diagnoses. *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 1–10, 2015.

[11] Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. CLPsych 2015 Shared Task: Depression and PTSD on Twitter. *the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 31–39, 2015.

[12] Glen Coppersmith, Craig Harman, and Mark Dredze. Measuring Post Traumatic Stress Disorder in Twitter. *In Proceedings of the 7th International AAAI Conference on Weblogs and Social Media (ICWSM)*., 2(1):23–45, 2014.

[13] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. Predicting Depression via Social Media. In *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media*, pages 128–138, 2013.

[14] Olive Jean Dunn. Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293):52–64, 1961.

[15] Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. Part-of-speech Tagging for Twitter: Annotation, Features, and Experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, pages 42–47, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[16] Su Golder, Shahd Ahmed, Gill Norman, and Andrew Booth. Attitudes toward the ethics of research using social media: A systematic review, June 2017.

[17] Sharath Chandra Guntuku, David B. Yaden, Margaret L. Kern, Lyle H. Ungar, and Johannes C. Eichstaedt. Detecting depression and mental illness on social media: an integrative review, 2017.

[18] Seda Gurses, Rebekah Overdorf, and Ero Balsa. POTs: The revolution will not be optimized? *11th Hot Topics in Privacy Enhancing Technologies (HotPETs)*, 2018.

[19] Harward Harvard Medical School. National Comorbidity Survey (NCS). https://www.hcp.med.harvard.edu/ncs/index.php, 2007. [Online; Accessed 26-April-2018, Ref Data Table 2: 12-month prevalence DSM-IV/WMH-CIDI disorders by sex and cohort (https://www.hcp.med.harvard.edu/ncs/ftpdir/table_ncsr_12monthprevgenderxage.pdf)].

[20] Melinda R Hess and Jeffrey D Kromrey. Robust confidence intervals for effect sizes: A comparative study of cohen'sd and cliff's delta under non-normality and heterogeneous variances. 2004.

[21] Alvarez-Conrad Jennifer, Zoellner Lori A., and Foa Edna B. Linguistic predictors of trauma pathology and physical health. *Applied Cognitive Psychology*, 15(7):S159–S170, 2001.

[22] Ron Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'95, pages 1137–1143, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc.

[23] Björn Lantz. The large sample size fallacy. *Scandinavian journal of caring sciences*, 27(2):487–492, 2013.

[24] Naomi Lee. Trouble on the radar. *The Lancet*, 384(29):1917, 2014.

[25] Huina Mao, Xin Shuai, and Apu Kapadia. Loose Tweets: An Analysis of Privacy Leaks on Twitter. In *Proceedings of the 10th Annual ACM Workshop on Privacy in the Electronic Society*, WPES '11, pages 1–12, New York, NY, USA, 2011. ACM.

[26] Jon D Mcauliffe and David M Blei. Supervised topic models. In *Advances in neural information processing systems*, pages 121–128, 2008.

[27] Jude Mikal, Samantha Hurst, and Mike Conway. Ethical issues in using Twitter for population-level depression monitoring: A qualitative study. *BMC Medical Ethics*, 17(1), 2016.

[28] Nilly Mor and Jennifer Winquist. Self-Focused Attention and Negative Affect : A Meta-Analysis. *Psychological bulletin*, 128(4):638–662, 2002.

[29] Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, and Nathan Schneider. Part-of-speech tagging for twitter: Word clusters and other advances. 2012.

[30] Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*, 2016.

[31] James W Pennebaker and Laura A King. Linguistic styles: Language use as an individual difference. *Journal of personality and social psychology*, 77(6):1296, 1999.

[32] Daniel Preoţiuc-Pietro, Johannes Eichstaedt, Gregory Park, Maarten Sap, Laura Smith, Victoria Tobolsky, H Andrew Schwartz, and Lyle Ungar. The role of personality, age, and gender in tweeting about mental illness. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 21–30, 2015.

[33] Philip Resnik, William Armstrong, Leonardo Claudino, Thang Nguyen, Viet-An Nguyen, and Jordan Boyd-graber. Beyond LDA : Exploring Supervised Topic Modeling for Depression-Related Language in Twitter. *Proceedings of the 52nd Workshop Computational Linguistics and Clinical Psychology*, 1(2014):99–107, 2015.

[34] Philip Resnik, William Armstrong, Leonardo Claudino, Thang Nguyen, Viet-An Nguyen, and Jordan Boyd-Graber. The University of Maryland CLPsych 2015 Shared Task System. *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 54–60, 2015.

[35] Philip Resnik, Anderson Garron, and Rebecca Resnik. Using Topic Modeling to Improve Prediction of Neuroticism and Depression. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1348–1353, 2013.

[36] Guillaume A. Rousselet. Cohen's d is biased. https://garstats.wordpress.com/2018/04/04/dbias/, 2018. [Online; Accessed 26-November-2018].

[37] Beatrice Santorini. Part-of-speech tagging guidelines for the Penn Treebank Project (3rd revision). *Technical Reports (CIS)*, page 570, 1990.

[38] H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin E.P. Seligman, and Lyle H Ungar. Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. *PLoS ONE*, 8(9), 2013.

[39] Manya Sleeper, Justin Cranshaw, Patrick Gage Kelley, Blase Ur, Alessandro Acquisti, Lorrie Faith Cranor, and Norman Sadeh. "I read my Twitter the next morning and was astonished" a conversational perspective on Twitter regrets. *Chi*, pages 3277–3286, 2013.

[40] Daria Smirnova, Paul Cumming, Elena Sloeva, Natalia Kuvshinova, Dmitry Romanov, and Gennadii Nosachev. Language patterns discriminate mild depression from normal sadness and euthymic state. *Frontiers in Psychiatry*, 9:105, 2018.

[41] Gail M Sullivan and Richard Feinn. Using effect size—or why the p value is not enough. *Journal of graduate medical education*, 4(3):279–282, 2012.

[42] Yla R Tausczik and James W Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54, 2010.

[43] The Federal Trade Comission. Data brokers: A call for transparency and accountability. *Data Brokers and the Need for Transparency and Accountability*, pages 1–101, 2014.

[44] Rianne Van der Zanden, Keshia Curie, Monique Van Londen, Jeannet Kramer, Gerard Steen, and Pim Cuijpers. Web-based depression treatment: Associations of clients' word use with adherence and outcome. *Journal of Affective Disorders*, 160:10–13, 2014.

[45] Yang Wang, Gregory Norcie, Saranga Komanduri, Alessandro Acquisti, Pedro Giovanni Leon, and Lorrie Faith Cranor. "I Regretted the Minute I Pressed Share": A Qualitative Study of Regrets on Facebook. In *Proceedings of the Seventh Symposium on Usable Privacy and Security*, SOUPS '11, pages 10:1–10:16, New York, NY, USA, 2011. ACM.

[46] WHO. WHO Depression Key Facts. http://www.who.int/en/news-room/fact-sheets/detail/depression, 2018. [Online; Accessed 26-April-2018].

[47] Johannes Zimmermann, Timo Brockmeyer, Matthias Hunn, Henning Schauenburg, and Markus Wolf. First-person Pronoun Use in Spoken Language as a Predictor of Future Depressive Symptoms: Preliminary Evidence from a Clinical Sample of Depressed Patients. *Clinical Psychology & Psychotherapy*, 24(2):384–391, mar 2017.

Appendix: Technical Details to Reproduce Experiments

This appendix explains the implementation level details such as the libraries and parameters that we used in our experiments. We have also made the code available at https://github.com/janithnw/twitter_mh_public. This appendix could be read as supplementary material for Section 3 and also as high-level details of the source code.

We used Python 3.6 with NumPy, Pandas and ScikitLearn for most of the experiments and Plotly to generate plots.

Preprocessing:

In the preprocessing stage, for all the tweets of each user we do the following:

- **Remove re-tweets and tweets with URLs**: We removed tweets in the dataset that were indicated as retweets and used a regular expression to detect URLs and removed tweets with URLs.
- **Handling Emojis**: We used pymoji library (https://pypi.org/project/pymoji/0.2.0/) to convert Unicode Emoji's to a string. For example 😂 is converted to: `_e_face_with_tears_of_joy_e` and this string is treated as a single token.
- **Tokenizing**: We used NLTK's TweetTokenizer (http://www.nltk.org/api/nltk.tokenize.html) which is a Twitter-aware tokenizer.
- **POS Tags**: We pre-computed the POS tags for each tweet and maintained them together with each tweet to be used in future steps. We used two POS Taggers: The NLTK POS Tagger (https://www.nltk.org/api/nltk.tag.html) and a Twitter specific POS Tagger from the TweetNLP project [15] (http://www.cs.cmu.edu/~ark/TweetNLP/#pos) and used the Python Wrapper available at https://github.com/ianozsvald/ark-tweet-nlp-python to invoke the tagger from python.

Computing Feature Sets:

We implemented each of the four featuresets as a class that implements Sklearn's `BaseEstimator` and `TransformerMixin` so that it exposes the `fit_transform` method, allowing them to be used in

Sklearn's `Pipeline` and `FeatureUnion` classes. This allowed us to combine different feature sets easily and use them with different classifiers and evaluation methods.

- **Bag-of-Words**: To compute the bag-of-words features we used Sklearn's `TfidfVectorizer` with `min_df: 0.1`.
- **Word-Clusters**: We use the 1000 hiarachical word clusters computed by Owoputi et al. [29] using 56 million English tweets (837 million tokens, available at http://www.cs.cmu.edu/~ark/TweetNLP/clusters/50mpaths2). The clusters specify a cluster identifier for each of the tokens in the dataset. We implemented a tokenizer that takes a string as input, tokenizes the string, and returns a list of cluster identifiers that appear in the string in the order they appear. This tokenizer is then used with a `TfidfVectorizer` to get the TF-IDF values for each cluster identifier.
- **POS Tags**: As mentioned earlier, we precomputed the two different POS tags for each tweet and stored them alongside each tweet. We implimented a tokenizer that takes a tweet as input and returns the POS tags as a list which is then used in a TF-IDF vectorizer to compute the feature vector.
- **Topic Models**: As mentioned in the paper we used the approach described by Resnik et al. [34]. Since we were not able to find the implementation of their approach we recreated this approach. We used LDAPlusPlus for both the LDA and sLDA computations. As described in Resnik et al.'s work, we used Pennebaker and King's [31] stream of consciousness essays to create priors for topics.
  We first computed a shared vocabulary that includes words from the stream of consciousness essays and our dataset. The vocabulary includes the union of all the words in the Twitter dataset that appears in more than 30 tweets and in less than 90% of the tweets (i.e. We set the following parameters on the SKlearn CountVectorizer `min_df=30` and `max_df=0.9`) and all the words in the stream of consciousness essays dataset that appear in more than 5 essays and in fewer than 90% of the essays (`min_df=5` and `max_df=0.9`). We used this shared vocabulary and the stream-of-consciousness essays to compute the input matrix for the LDA step to learn topic priors. We used the following parameters: `topics:50, iterations:200 e_step_iterations:200, e_step_tolerance:0.1`. LDAPlusPlus allows this trained model to be used as a prior in the sLDA step. To train the sLDA model, we computed the input matrix using the shared vocabulary and the weekly aggregated user tweets. We used the following parameters for the sLDA step: `topics:50, iterations:100 e_step_iterations:100, e_step_tolerance:0.1`. The labels of each user were considered as the label for each set of weekly-aggregated tweets. When computing the feature vector for each user, we weighted features based on the fraction of the user's tweets associated with each week.

Removing Direct Mentions of Mental Health Content

Tweet Labelling Protocol

The following labelling protocol was agreed upon and used by two of the authors after evaluating a sample set of tweets.

Depression vs Control (DvC) Dataset:

- **Depression**: A non satirical tweet:
  - That mentions the words depressed, depression, or depress
  - Tweet that implies user has depression. For example, use of anti-depressants
- **Mental Health**: A non satirical tweet
  - That mentions a mental illness other than depression (anxiety, bipolar disorder, PTSD, etc)
  - Tweets about self-harm, suicidal thoughts, anorexia
  - Tweets about mental health advocacy (Tweets that include hashtags like #sicknotweek)
- **Other**: A tweet that does not belong to any of the above categories

PTSD vs Control (PvC) Dataset:

- **PTSD**: A non satirical tweet that mentions the words PTSD, P.T.S.D, Post Traumatic Stress, Stress Disorder, Post Traumatic Syndrome, or a similar phrase
- **Mental Health**: A non satirical tweet
  - That mentions a mental illness other than PTSD (anxiety, bipolar disorder, depression, etc)
  - Tweets about self-harm, suicidal thoughts, anorexia
  - Tweets about mental health advocacy (Tweets that include hashtags like #sicknotweek)

– **Other**: A tweet that does not belong to any of the above categories

Classifier details

As discussed in Section 3.4, we used a classifier to identify tweets that mention mental health-related issues. This classifier contained three feature sets.

– **Bag-of-Words**: We used `TfidfVectorizer` with the following parameters: `min_df:30, max_df:0.9`
– **Word Clusters**: We used the same word cluster implementation as above with the following parameters: `min_df:10, max_df:0.9`
– **Custom Word Lists**: We implemented a feature extractor that would return 1 if any of the given phrases are included in a tweet. We used this extractor with the following 4 different world lists: [anti-depressant, mentalillness, mental illness, brain disease, mental health, depressive disorder, mental disorder, suicidal, suicide, anxiety, depression, bipolar, schizophrenia], [suicidal, suicide, self-hate], [self-harm, self harm], [ptsd, p.t.s.d, post-traumatic, post traumatic, stress disorder]

We used a Sklearn's Random Forest Classifier with the number of estimators (`n_estimators`) set to 500.

Evaluating Prediction Accuracy

We used SKLearn's `LinearSVM` classifier with the default parameters wrapped in SKLearn's `CalibratedClassifierCV` with cross-validation folds set to 3, which outputs a probability value for predictions. To evaluate the performance of each model we ran 10-fold cross-validation using SKLearn's `model_selection.cross_validate` method.

Feature Analysis

We used three measures to evaluate feature importance. We used Scipy `ttest_ind` method to compute the p-value. Cohen's d measure and information gain were implemented according to the definitions given in Section 3.5. We used the Python WordCloud library (https://amueller.github.io/word_cloud/) to create the word clouds. To generate the word clusters feature analysis results (Figure 3), we first generated the individual word clouds with appropriate sizes, and used the image editing tool (Gimp) to combine them to form one image.

Other Implementation Details

The Misclassifiction Analysis (Section 3.6) and Mitigations (Section 4.4) both use the tweet $\texttt{impact}(M, t_i)$ function. We implemented this as a python function that takes as input a pre-trained Sklearn `pipeline` and an individual user's tweets and outputs a list of tweet-impact values for each tweet. The experiments for the Mitigations section were implemented as follows: for a given pre-trained Sklearn `pipeline` and an individual user, we add, remove, or replace one tweet and run the prediction step again. We repeat this process until the classifier prediction changes. When removing tweets we remove tweets in the descending order of each tweet's impact value (i.e: remove most positive-leaning tweet first). When adding tweets, we add the most negative-leaning tweets from the control group. The timestamp of the newly added tweets were set so that the newly added tweets were appended as one tweet per day.