Quantifying Mental Health Signals in Twitter

Glen Coppersmith **Mark Dredze** **Craig Harman**
Human Language Technology Center of Excellence
Johns Hopkins University
Balitmore, MD, USA

Abstract

The ubiquity of social media provides a rich opportunity to enhance the data available to mental health clinicians and researchers, enabling a better-informed and better-equipped mental health field. We present analysis of mental health phenomena in publicly available Twitter data, demonstrating how rigorous application of simple natural language processing methods can yield insight into specific disorders as well as mental health writ large, along with evidence that as-of-yet undiscovered linguistic signals relevant to mental health exist in social media. We present a novel method for gathering data for a range of mental illnesses quickly and cheaply, then focus on analysis of four in particular: post-traumatic stress disorder (PTSD), depression, bipolar disorder, and seasonal affective disorder (SAD). We intend for these proof-of-concept results to inform the necessary ethical discussion regarding the balance between the utility of such data and the privacy of mental health related information.

1 Introduction

While mental health issues pose a significant health burden on the general public, mental health research lacks the quantifiable data available to many physical health disciplines. This is partly due to the complexity of the underlying causes of mental illness and partly due to longstanding societal stigma making the subject all but taboo. Lack of data has hampered mental health research in terms of developing reliable diagnoses and effective treatment for many disorders. Moreover, population-level analysis via traditional methods is time consuming, expensive, and often comes with a significant delay.

In contrast, social media is plentiful and has enabled diverse research on a wide range of topics, including political science (Boydstun et al., 2013), social science (Al Zamal et al., 2012), and health at an individual and population level (Paul and Dredze, 2011; Dredze, 2012; Aramaki et al., 2011; Hawn, 2009). Of the numerous health topics for which social media has been considered, mental health may actually be the most appropriate. A major component of mental health research requires the study of behavior, which may be manifest in how an individual acts, how they communicate, what activities they engage in and how they interact with the world around them including friends and family. Additionally, capturing population level behavioral trends from Web data has previously provided revolutionary capabilities to health researchers (Ayers et al., 2014). Thus, social media seems like a perfect fit for studying mental health in both individual and overall trends in the population. Such topics have already been the focus of several studies (Coppersmith et al., 2014; De Choudhury et al., 2014; De Choudhury et al., 2013d; De Choudhury et al., 2013b; De Choudhury et al., 2013c; Ayers et al., 2013).

What can we expect to learn about mental health by studying social media? How does a service like Twitter inform our knowledge in this area? Numerous studies indicate that language use, social expression and interaction are telling indicators of mental health. The well-known Linguistic Inquiry Word Count (LIWC), a validated tool for the psychometric analysis of language data (Pennebaker et al., 2007), has been repeatedly used to study language associated with all types of disorders (Resnik et al., 2013; Alvarez-Conrad et al., 2001; Tausczik and Pennebaker, 2010). Furthermore, social media is by nature *social*, which means that social patterns, a critical part of mental health and illness, may be readily observable in raw Twitter data. Thus, Twitter and other social media provide a unique quantifiable perspective on human behavior that may otherwise go unobserved, suggesting it as a powerful tool for mental health researchers.

The main vehicle for studying mental health in social media has been the use of surveys, e.g., depression battery (De Choudhury, 2013) or personality test (Schwartz et al., 2013), to determine characteristics of a user coupled with analyzing their corresponding social media data. Work in this area has mostly focused on depression (De Choudhury et al., 2013d; De Choudhury et al., 2013b; De Choudhury et al., 2013c), and the number of users is limited by those that can complete the appropriate survey. For example, De Choudhury et al. (2013d) solicited Twitter users to take the CES-D and to share their public Twitter profile, analyzing linguistic and behavioral patterns. While this type of study has produced high quality data, it is limited in size (by survey respondents) and scope (to diagnoses which have a battery amenable to administration over the internet).

In this paper we examine a range of mental health disorders using *automatically derived* samples from large amounts of Twitter data. Rather than rely on surveys, we automatically identify self-expressions of mental illness diagnoses and leverage these messages to construct a labeled data set for analysis. Using this dataset, we make the following contributions:

- We demonstrate the effectiveness of our automatically derived data by showing that statistical classifiers can differentiate users with four different mental health disorders: depression, bipolar, post traumatic stress disorder and seasonal affective disorder.

- We conduct a LIWC analysis of each disorder to measure deviations in each illness group from a control group, replicating previous findings for depression and providing new findings for bipolar, PTSD and SAD.

- We conduct an open-vocabulary analysis that captures language use relevant to mental health beyond what is captured with LIWC.

Our results open the door to a range of large scale analysis of mental health issues using Twitter.

2 Related Work

For a good retrospective and prospective summary of the role of social media in mental health research, we refer the reader to De Choudhury (2013). De Choudhury identifies ways in which NLP has and can be used on social media data to produce what the relevant mental health literature would predict, both at an individual level and a population level. She proceeds to identify ways in which these types of analyses can be used in the near and far term to influence mental health research and interventions alike.

Differences in language use have been observed in the personal writing of students who score highly on depression scales (Rude et al., 2004), forum posts for depression (Ramirez-Esparza et al., 2008), self narratives for PTSD (He et al., 2012; D'Andrea et al., 2011; Alvarez-Conrad et al., 2001), and chat rooms for bipolar (Kramer et al., 2004). Specifically in social media, differences have previously been observed between depressed and control groups (as assessed by internet-administered batteries) via LIWC: depressed users more frequently use first person pronouns (Chung and Pennebaker, 2007) and more frequently use negative emotion words and anger words on Twitter, but show no differences in positive emotion word usage (Park et al., 2012). Similarly, an increase in negative emotion and first person pronouns, and a decrease in third person pronouns, (via LIWC) is observed, as well as many manifestations of literature findings in the pattern of life of depressed users (e.g., social engagement, demographics) (De Choudhury et al., 2013d). Differences in language use in social media via LIWC have also been observed between PTSD and control groups (Coppersmith et al., 2014).

For population-level analysis, surveys such as the Behavioral Risk Factor Surveillance System (BRFSS) are conducted via telephone (Centers for Disease Control and Prevention (CDC), 2010). Some of these surveys cover relatively few participants (often in the thousands), have significant cost, and have long delays between data collection and dissemination of the findings. However, De Choudhury et al. (2013c) presents a promising population-level analysis of depression that highlights the role of NLP and social media.

3 Data

All data we obtain is public, posted between 2008 and 2013, and made available from Twitter via their application programming interface (API). Specifically, this does **not** include any data that has

Genuine Statements of Diagnosis
In loving memory my mom, she was only 42, I was 17 & taken away from me. I was diagnosed with having P.T.S.D LINK
So today I started therapy, she diagnosed me with anorexia, depression, anxiety disorder, post traumatic stress disorder and wants me to
@USER The VA diagnosed me with PTSD, so I can't go in that direction anymore
I wanted to share some things that have been helping me heal lately. I was diagnosed with severe complex PTSD and... LINK
Disingenuous Statements of Diagnosis
"I think I'm I'm diagnosed with SAD. Sexually active disorder" -anonymous
LOL omg my bro the "psychologist" just diagnosed me with seasonal ADHD AHAHAHAAAAAAAAAAA IM DYING.
The winter blues: Yesterday I was diagnosed with seasonal affective disorder. Now, this sounds a lot more dramat... LINK

Table 1: Examples found via regular expression keyword search for diagnosis tweets.

been marked as 'private' by the author or any direct messages.

Diagnosed Group We seek users who publicly state that they have been diagnosed with various mental illnesses. Users may make such a statement to seek support from others in their social network, to fight the taboo of mental illness, or perhaps as an explanation of some of their behavior. Tweets were obtained using regular expressions on a large multi-year health related collection, e.g. "I was diagnosed with X." We searched for four conditions: depression, bipolar disorder, post traumatic stress disorder (PTSD) and seasonal affective disorder (SAD). The matched diagnosis tweets were manually labeled as to whether the tweet contained a genuine statement of a mental health diagnosis. Table 1 shows examples of both genuine statements of diagnosis and disingenuous statements (often jokes or quotes).

Next, we retrieved the most recent tweets (up to 3200) for each user with a genuine diagnosis tweet. We then filtered the users to remove those with fewer than 25 tweets and those whose tweets were not at least 75% in English (measured using the Compact Language Detector[1]). These filtering steps left us with users that were considered positive examples. Table 2 indicates the number of users and tweets found for each of the mental health categories examined. We manually examined and annotated only half the diagnosis statements for depression – indicating there are likely 800-900 depression users available via these automatic methods from our collection, compared to the 117 obtained via the methods of De Choudhury et al. (2013d). Additionally, we emphasize the low cost and effort of our automated effort as compared to their crowdsourced survey methods. The difference in collection methods also suggests that the two have a reasonable chance of being complementary. This is especially significant when considering disorders with lower incidence rates than depression (arguably the highest), where respondents to crowdsourced surveys or self-stated diagnoses alike are rare.

This method is similar in spirit to that of De Choudhury et al. (2013c), where they inferred a tweet-level classifier for depression from user-level labels (specifically, tweets from the past three months from users scoring highly on CES-D for the positive class and conversely for the negative).

Control Group To build models for analysis and to validate the data, we also need a sample of the general population to use as an approximation of community controls. We follow a similar process: randomly select 10k usernames from a list of Twitter users who posted to a separate random historical collection within a selected two week window, downloaded the 3200 most recent tweets from these users, and apply our two filters: at least 25 tweets and 75% English. This yields a control group of 5728 random users, whose 13.7 million tweets were used as negative examples.

Caveats Our method for finding users with mental health diagnoses has significant caveats: **1)** the method may only capture a subpopulation of each disorder (i.e., those who are speaking publicly about what is usually a very private matter), which may not truly represent all aspects of the population as a whole. **2)** This method in no way verifies whether this diagnosis is genuine (i.e., people are not always truthful in self-reports). However, given the stigma often associated with mental illness, it seems unlikely users would tweet that they are diagnosed with a condition they do not have. **3)** The control group is likely contami-

[1] https://code.google.com/p/cld2/

	Match	Users	Tweets
Bipolar	6k	394	992k
Depression	5k	441	1.0m
PTSD	477	244	573k
SAD	389	159	421k
Control	10k	5728	13.7m

Table 2: Number of users **match**ing the diagnosis regular expression, **users** labeled with genuine diagnoses and **tweets** retrieved from diagnosed users for each mental health condition.

nated by the presence of users that are diagnosed with the various conditions investigated. We make no attempt to remove these users, and if we assume that the prevalence of each disorder in the general population is similar in our control groups, we likely have hundreds of such diagnosed users contaminating our control training data. **4)** Twitter users are not an entirely representative sample of the population as a whole. Despite these caveats, we find that this method yielded promising results as discussed in the next sections.

Comorbidity Since some of these disorders have high comorbidity, there are some users in more than one class (e.g., those that state a diagnosis for PTSD and depression): Bipolar and depression have 19 users in common (4.8% of the bipolar users, 4.3% of the depression users), PTSD and depression share 10 (4.0% of PTSD, 2.2% of depression), and bipolar and PTSD share 9 (2.2% of bipolar, 3.6% of PTSD). Two users state diagnosis of bipolar, PTSD and depression (less than 1% of each set). No users stated diagnoses of both SAD and any other condition investigated.

4 Methods

We quantify various aspects of each user's language usage and pattern of life via automated methods, extracting features for subsequent machine learning. We use these to (1) replicate previous findings, (2) build classifiers to separate diagnosed from control users, and (3) introspect on those classifiers. Introspection here shows us what quantified signals in the content the classifiers base their decision on, and thus we can gain intuition about what signals are present in the content relevant to mental health.

4.1 Linguistic Inquiry Word Count (LIWC)

LIWC provides clinicians with a tool for gathering quantitative data regarding the state of a patient from the patient's writing (Pennebaker et al., 2007). Previous work has found signal in the 'positive affect' and 'negative affect' categories of the LIWC when applied to social media (including Twitter), so we examine their correlations separately, as well as in the context of other LIWC categories (De Choudhury et al., 2013a). In all, we examine some of the LIWC categories directly (*Swear*, *Anger*, *PosEmo*, *NegEmo*, *Anx*) and combine pronoun classes by linguistic form: *I* and *We* classes are combined to form *Pro1*, *You* becomes *Pro2* and *SheHe* and *They* become *Pro3*. Each of these classes provides one feature used by subsequent machine learning and our other analyses.

4.2 Language Models (LMs)

Language models are commonly used to estimate how likely a given sequence of words is. Generally, an n-gram language model refers to a model that examines strings of up to n words long. This is less than ideal for applications in social media: spelling errors, shortenings, space removal, and other aspects of social media data (especially Twitter) confounds many traditional word-based approaches. Thus, we employ two LMs, first a traditional 1-gram LM (ULM) that examines the probability of each whole word. Second, a character 5-gram LM (CLM) to examine sequences of up to 5 characters.

LMs model the likelihood of sequences from training data. In our case, we build one of each model from the positive class (tweets from one class of diagnosed users – e.g., PTSD), yielding ULM$^+$ and CLM$^+$. We also build one of each model from the negative class (control users), yielding ULM$^-$ and CLM$^-$. We score each tweet by computing these probabilities and classifying it according to which model has a higher probability (e.g., for a given tweet, is ULM$^+ >$ ULM$^-$?).

4.3 Pattern of Life Analytics

For brevity, we only briefly discuss the pattern of life analytics, since they do not depend on significant NLP. They examine how correlates found to be significant in the mental health literature may manifest and be measured in social media data. These are all imperfect proxies for the findings from the literature, but our experiments will demonstrate that they do collectively provide information relevant to mental health.

For each of the following analytics we extract one feature to use in subsequent machine learning. Social engagement has been correlated with

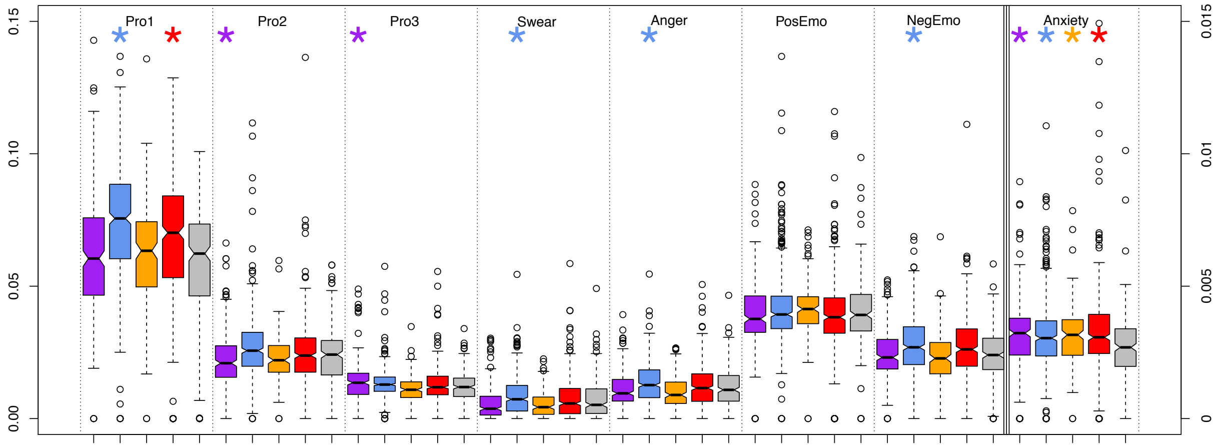


Figure 1: Box and whiskers plot of proportion of tweets each user has (y-axis) matching various LIWC categories. Each bar represents one LIWC category for one condition – PTSD in purple, depression in blue, SAD in orange, bipolar in red and control in gray. *Anxiety* occurs an order of magnitude less often than the others, so its proportion is on the right y-axis (and thus not comparable to the others). Statistically significant deviations from control users are denoted by asterisks.

positive mental health outcomes (Greetham et al., 2011; Berkman et al., 2000; Organization, 2001; De Choudhury et al., 2013d), which is difficult to measure directly so we examine various ways in which this may be manifest in a user's tweet stream: *Tweet rate* measures how often a twitter user posts (a measure of overall engagement with this social media platform) and *Proportion of tweets with @mentions* measures how often a user posts 'in conversation' (for lack of better terms) with other users. *Number of @mentions* is a measure of how often the user in question engages other users, while *Number of self @mentions* is a measure of how often the user responds to mentions of themselves (since users rarely include their own username in a tweet). To estimate the size of a user's social network, we calculate *Number of unique users @mentioned* and *Number of users @mentioned at least 3 times*, respectively.

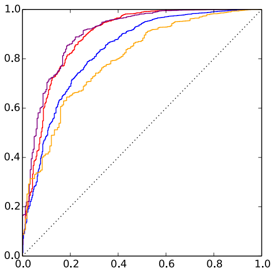
For each of the following analytics, we calculate the proportion of a user's tweets that the analytic finds evidence in: *Insomnia* and sleep disturbance is often a symptom of mental health disorders (Weissman et al., 1996; De Choudhury et al., 2013d), so we calculate the proportion of tweets that a user makes between midnight and 4am according to their local timezone. *Exercise* has also been correlated with positive mental health outcomes (Penedo and Dahn, 2005; Callaghan, 2004), so we examine tweets mentioning one of a small set of exercise-related terms. We also use an English *sentiment* analysis lexicon from Mitchell et al. (2013) to score individual tweets according to the presence and valence of sentiment words. We apply no thresholds, so any tweet with a sentiment score above 0 was considered *positive*, below 0 was considered *negative*, and those with score 0 were considered to have *no sentiment*. Thus we use the proportion of *Insomnia*, *Exercise*, *Positive Sentiment* and *Negative Sentiment* tweets as features in subsequent machine learning and analysis.

5 Results

We present three types of experiments to evaluate the quality and character of these data, and to demonstrate some quantifiable mental health signals in Twitter. First, we validate our method for obtaining data by replicating previous findings using LIWC. Next, we build classifiers to distinguish each group from the control group, demonstrating that there is useful signal in the language of each group, and compare these classifiers. Finally, we analyze the correlations between our analytics and classifiers to uncover relationships between them and derive insight into quantifiable and relevant mental health signals in Twitter.

Validation First, we provide some validation for our novel method for gathering samples. We demonstrate that language use, as measured by LIWC, is statistically significantly different between control and diagnosed users. Figure 1 shows the proportion of tweets from each user that scores positively on various LIWC categories (i.e., have at least one word from that category). Box-and-whiskers plots (Tukey, 1977)[2] summarize a distribution of observations and ease com-

[2] For a modern implementation see Wickham (2009).



False Alarm:	0.1	0.2
Bipolar	0.64	0.82
Depression	0.48	0.68
PTSD	0.67	0.81
SAD	0.42	0.65

Figure 2: ROC curves for separating diagnosed from control users, compared across disorders: bipolar in red, depression in blue, PTSD in purple, SAD in orange. The precision (diagnosed, correctly labeled) for each disorder at false alarm (control, labeled as diagnosed) rates of 10% and 20% are shown to the right of the ROC curve. Chance performance is indicated by the dotted black line.

parison between them (here, each observation is the proportion of a user's tweets that score positively on LIWC). The median of the distribution is the black horizontal line in the middle of the bar, the bar covers the inter quartile range (where 50% of the observations lie), the whiskers are a robust estimate of the extent of the data, with outliers plotted as circles beyond the whiskers. An approximation of statistical significance is indicated by the pinched in notches on each bar. If the notches on the bars do not overlap, the differences between those distributions is different (α<0.05, 95% confidence interval). Each bar is colored according to diagnosis, and each group of 5 bars notes the scores for one LIWC category. Differences that reach statistical significance from the control group are noted with asterisks (e.g., *Pro1*, *Swear*, *Anger*, *NegEmo* and *Anxiety* are statistically significantly different for the depression group). Importantly, this replicates previous findings of significant differences between depressed users (according to an internet-administered diagnostic battery): significant increases are expected in *NegEmo*, *Anger*, *Pro1* and *Pro3* and no change in *PosEmo*, given all previous work (Park et al., 2012; Chung and Pennebaker, 2007; De Choudhury et al., 2013d). We replicate all these findings except the increase in *Pro3* (which only De Choudhury et al. (2013d) found), which validates our data collection methods.

Classification We next explore the ability of the various analytics to separate diagnosed from control users and assess performance on a leave-one-out cross-validation task. We train a log linear classifier on the features described in §4 using scikit-learn (Pedregosa et al., 2011).

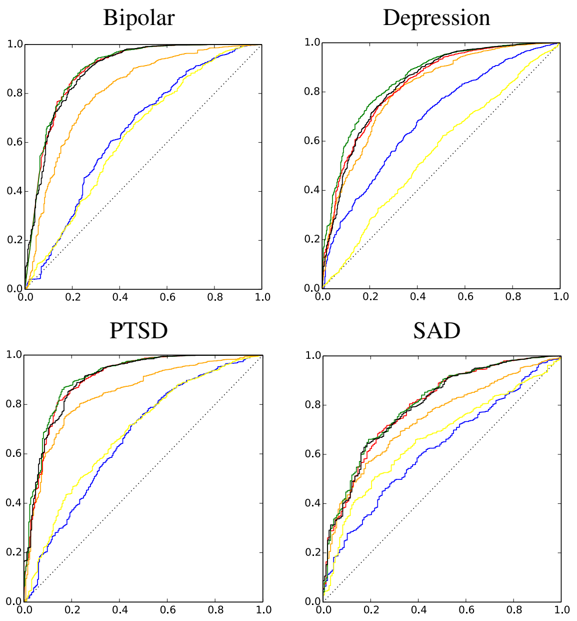


Figure 3: ROC curves of performance of individual analytics for each disorder: LIWC in blue, pattern of life in yellow, CLM in red, ULM in green, all in black. Chance performance is indicated by the dotted black line.

The receiver operating characteristic (ROC) curves in Figures 2 and 3 demonstrate performance of the various classifiers at the task of separating diagnosed from control groups. In all cases, the correct detections (or hits) are on the y-axis and the false detections (or false alarms) are on the x-axis. Figure 2 compares performance across diagnoses, one line per disorder.

Figure 3 shows one plot per mental health condition, with the performance of the various analytics, individually and in concert as individual ROC curves. A few trends emerge – **1)** All analytics show some ability to separate the classes, indicating they are finding useful signals. **2)** The LMs provide superior performance to the other analytics, indicating there are more signals present in the language than are captured by LIWC and pattern-of-life analytics. For readability we do not show the performance of all combinations of analytics, but they perform as expected: any set of them perform equal to or better than their individual components. Taken together, this indicates that there is information relevant to separating diagnosed users from controls in all the analytics discussed here. Furthermore, this highlights that there remains significant signals to be uncovered and understood in the language of social media.

These trends also allow us to compare the disorders as manifest in language usage, though this tends to raise more questions than it answers. Generally, the pattern-of-life analytics and LIWC are on par, but this is decidedly not true for depression, where pattern-of-life seems to perform especially poorly, and for SAD, where pattern-of-life seems to perform especially well. This indicates that the depression users have patterns-of-life that look more similar to the controls than is the case for the other disorders (perhaps especially surprising given the inclusion of the sentiment lexicon) and that there may be significant correlation between pattern-of-life factors and SAD.

5.1 Analytic Introspection

To examine correlations between the analytics and the linguistic content they depend on, we scored a random subset of 1 million tweets from control users with each of the linguistic analytics, and plot their Pearson's correlation coefficients (r) in Figure 4. A simple overlap of wordlists is not sufficient to assess the true utility of these methods since it does not take into account the frequency of occurrence of each word, nor the correlation between these words in real data (e.g., does a classifier based on the LIWC category *Swear* provide redundant information to the sentiment analysis). Each row and column in Figure 4 represents one of the 17 analytics, in the same order. Colors denote Bonferroni-corrected Pearson's r for statistically significant correlations between the analytic on the row and column. Correlations that do not reach statistical significance are in aquamarine (corresponding to r=0). Excluded for brevity is a sanity check of a χ^2 test between the analytics to assert they were scoring significantly differently.

The strong correlations between the various LIWC analytics, notably *Swear*, *Anger* and *NegEmo*, likely indicates that the analytics are triggered by the same word(s) – in this case profanity. Similarly for LIWC's *PosEmo* and the sentiment lexicon – 'happy' for example. The correlation between CLM for various diagnoses is particularly intriguingly, as it is in line with known patterns of comorbidity: major depressive disorder, PTSD, and bipolar all have observed comorbidity (Brady et al., 2000; Campbell et al., 2007; McElroy et al., 2001) while SAD is currently considered a specifier of major depressive disorder or bipolar disorder (American Psychiatric Association, 2013; Lurie et al., 2006), without published findings indicating comorbidity. Indeed our small sample dataset follows the same trends, where we observed users with multiple diagnoses exist within depression, PTSD, and bipolar, but none exist with SAD. The correlation observed is too large to be solely attributed to those users shared between the groups, though (correlations at most $r = 0.05$ would be attributable to that alone). Furthermore, when taken in combination with the different patterns exhibited by the groups as seen in Figure 1, this correlation is not solely attributable to LIWC categories either. At its core, these correlations seem to suggest that similar language is employed by users diagnosed with these occasionally comorbid disorders, and dissimilar language by users with SAD. This should be taken as merely suggestive of the type of analysis one could do, though, since the literature does not present a strong and clear prediction for the comorbidity and exhibited symptoms (to include language use).

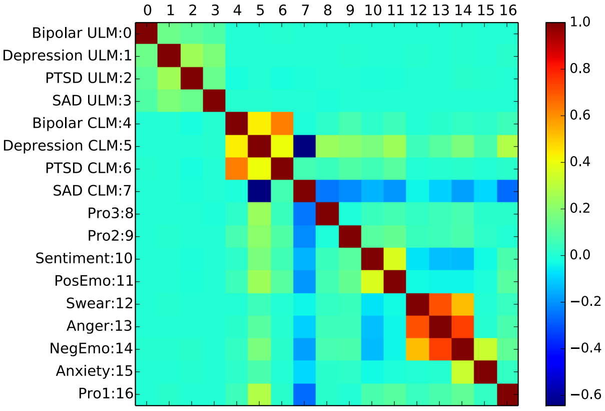


Figure 4: Pearson's r correlations between various analytics, color indicates the strength of statistically significant correlations, or 0 (aquamarine) otherwise. Bonferroni corrected, each comparison is significant only if α<0.0002). Rows and columns represent the analytics in the same order, so the diagonal is self-correlation.

Interestingly, the lack of (or negative) correlation between most of the analytics again highlights the complexity of the mental illnesses and the divergent signals it presents. Additionally, the lack of correlation between ULM and the other models is to be expected, since they are basing their scores on significantly more words (or different signals as is the case for CLM). Each one of these analytics is highly imperfect, and often give contradictory evidence, but when combined, the machine learning algorithms are able to sort through the conflicting signals with some success.

Analytic	Example Tweet Text
Bipolar LM	I'm insecure because being around your ex of 4 years little sister, makes me feel a slight bit uncomfortable. Ok.
Depression LM	Pain has a weird way of working. You're still the same person from before the pain, but that person is underneath & doesn't come out.
PTSD LM	Don't wanna get out my bed but I really need to get up & prepare myself for work
Sentiment(+)	NAME is absolutely unbelievable, he just gets better and better every time I see him. The best play in the world, no doubt about it.
Sentiment(-)	I hate losing people in my life. I try so hard to not let it happen
PosEmo	Wowee...that was a hectic day... Got more done than expected but so glad to be in bed now. Grateful for my supportive husband & loving pooch
Functioning	if i had a dollar for all the grammatical errors ive ever typed, my college tuition, book cost, and dorm rent would be paid in full
NegEmo	My tooth hurts, my neck hurts, my mouth hurts, my toungue hurts, my head hurts...kill me now.
Anx	don't stress over someone who is going to stress over you..
Anger	Ugly n arrogant sums everytin up.shdnt hv ffd her seff

Table 3: Example high scoring tweets from each analytic.

6 Conclusion

We demonstrate quantifiable signals in Twitter data relevant to bipolar disorder, major depressive disorder, post-traumatic-stress disorder and seasonal affective disorder. We introduce a novel method for automatic data collection and validate its veracity by **1)** replicating observations of significant differences between depressed and control user groups and **2)** constructing classifiers capable of separating diagnosed from control users for each disorder. This data allows us to demonstrate equivalent differences in language use (according to LIWC) for bipolar, PTSD, and SAD. Furthermore, we provide evidence that more information relevant to mental health is encoded in language use in social media (above and beyond that captured by methods based on the mental health literature). By examining correlations between the various analytics investigated, we provide some insight into what quantifiable linguistic information is captured by our classifiers. We finally demonstrate the utility of examining multiple disorders simultaneously and other larger analyses, difficult or impossible with other methods.

Crucially, we expect that these novel data collection methods can provide complementary information to existing survey-based methods, rather than supplant them. For many disorders rarer than depression (which has comparatively high incidence rates), we suspect that finding any data will be a challenge, in which case combining these methods with the existing survey collection methods may be the best way to obtain sufficient amounts of data for statistical analyses.

Since the LMs take more information into account when modeling the language usage of diagnosed and control users, it is unsurprising that they outperform LIWC and pattern-of-life analyses alone, but this is evidence of as-of-yet undiscovered linguistic differences between diagnosed and control users for all disorders investigated. Uncovering and interpreting these signals can be best accomplished through collaboration between NLP and mental health researchers.

Naturally, some caveats come with these results: while identifying genuine self-statements of diagnosis in Twitter works well for some conditions, others exist for which there were few or no diagnoses stated. For Alzheimer's, the demographic with the majority of diagnoses does not frequently use Twitter (or likely any social media). Eating disorders are also elusive via this method, though related automatic methods (e.g., using disorder-related hashtags) may address this. Finally, those willing to publicly reveal a mental health diagnosis may not be representative of the population suffering from that mental illness.

All these experiments, taken together, indicate that there are a diverse set of quantifiable signals relevant to mental health observable in Twitter. They indicate that individual- and population-level analyses can be made cheaper and more timely than current methods, yet there remains as-of-yet untapped information encoded in language use – promising a rich collaboration between the fields of natural language processing and mental health.

Acknowledgments: The authors would like to thank Kristy Hollingshead for thoughtful comments and contributions throughout this research.

References

Faiyaz Al Zamal, Wendy Liu, and Derek Ruths. 2012. Homophily and latent attribute inference: Inferring latent attributes of Twitter users from neighbors. In *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM)*.

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

American Psychiatric Association. 2013. *Diagnostic Statistical Manual 5*. American Psychiatric Association.

Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. 2011. Twitter catches the flu: Detecting influenza epidemics using twitter. In *Empirical Natural Language Processing Conference (EMNLP)*.

John W. Ayers, Benjamin M. Althouse, Jon-Patrick Allem, J. Niels Rosenquist, and Daniel E. Ford. 2013. Seasonality in seeking mental health information on google. *American journal of preventive medicine*, 44(5):520–525.

John W. Ayers, Benjamin M. Althouse, and Mark Dredze. 2014. Could behavioral medicine lead the web data revolution? *Journal of the American Medical Association (JAMA)*, February 27.

Lisa F. Berkman, Thomas Glass, Ian Brissette, and Teresa E. Seeman. 2000. From social integration to health: Durkheim in the new millennium? *Social Science & Medicine*, 51(6):843–857, September.

Amber Boydstun, Rebecca Glazier, Timothy Jurka, and Matthew Pietryka. 2013. Examining debate effects in real time: A report of the 2012 React Labs: Educate study. *The Political Communication Report*, 23(1), February. [Online; accessed 25-February-2014].

Kathleen T. Brady, Therese K. Killeen, Tim Brewerton, and Sylvia Lucerini. 2000. Comorbidity of psychiatric disorders and posttraumatic stress disorder. *Journal of Clinical Psychiatry*.

Patrick Callaghan. 2004. Exercise: a neglected intervention in mental health care? *Journal of Psychiatric and Mental Health Nursing*, 11:476–483.

Duncan G. Campbell, Bradford L. Felker, Chuan-Fen Liu, Elizabeth M. Yano, JoAnn E. Kirchner, Domin Chan, Lisa V. Rubenstein, and Edmund F. Chaney. 2007. Prevalence of depression-PTSD comorbidity: Implications for clinical practice guidelines and primary care-based interventions. *Journal of General Internal Medicine*, 22(6):711–718.

Centers for Disease Control and Prevention (CDC). 2010. Behavioral risk factor surveillance system survey data.

Cindy Chung and James Pennebaker. 2007. The psychological functions of function words. *Social communication*, pages 343–359.

Glen A. Coppersmith, Craig T. Harman, and Mark Dredze. 2014. Measuring post traumatic stress disorder in Twitter. In *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM)*.

Wendy D'Andrea, Pearl H. Chiu, Brooks R. Casas, and Patricia Deldin. 2011. Linguistic predictors of post-traumatic stress disorder symptoms following 11 September 2001. *Applied Cognitive Psychology*, 26(2):316–323, October.

Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013a. Major life changes and behavioral markers in social media: Case of childbirth. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW)*.

Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013b. Predicting postpartum changes in emotion and behavior via social media. In *Proceedings of the ACM Annual Conference on Human Factors in Computing Systems (CHI)*, pages 3267–3276. ACM.

Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013c. Social media as a measurement tool of depression in populations. In *Proceedings of the Annual ACM Web Science Conference*.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013d. Predicting depression via social media. In *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM)*.

Munmun De Choudhury, Andres Monroy-Hernandez, and Gloria Mark. 2014. " narco" emotions: Affect and desensitization in social media during the mexican drug war.

Munmun De Choudhury. 2013. Role of social media in tackling challenges in mental health. In *Proceedings of the 2nd International Workshop on Socially-Aware Multimedia*, pages 49–52.

Mark Dredze. 2012. How social media will change public health. *IEEE Intelligent Systems*, 27(4):81–84.

Danica Vukadinovic Greetham, Robert Hurling, Gabrielle Osborne, and Alex Linley. 2011. Social networks and positive and negative affect. *Procedia - Social and Behavioral Sciences*, 22:4–13, January.

Carleen Hawn. 2009. Take Two Aspirin And Tweet Me In The Morning: How Twitter, Facebook, And Other Social Media Are Reshaping Health Care. *Health Affairs*, 28(2):361–368.

Qiwei He, Bernard P. Veldkamp, and Theo de Vries. 2012. Screening for posttraumatic stress disorder using verbal features in self narratives: A text mining approach. *Psychiatry Research*.

Adam D. I. Kramer, Susan R. Fussell, and Leslie D. Setlock. 2004. Text analysis as a tool for analyzing conversation in online support groups. In *Proceedings of the ACM Annual Conference on Human Factors in Computing Systems (CHI)*.

Stephen J. Lurie, Barbara Gawinski, Deborah Pierce, and Sally J. Rousseau. 2006. Seasonal affective disorder. *American family physician*, 74(9).

Susan L. McElroy, Lori L. Altshuler, Trisha Suppes, Paul E. Keck, Mark A. Frye, Kirk D. Denicoff, Willem A. Nolen, Ralph W. Kupka, Gabriele S. Leverich, Jennifer R. Rochussen, A. John Rush Rush, and Robert M. Post Post. 2001. Axis I psychiatric comorbidity and its relationship to historical illness variables in 288 patients with bipolar disorder. *American Journal of Psychiatry*, 158(3):420–426.

Margaret Mitchell, Jacqueline Aguilar, Theresa Wilson, and Benjamin Van Durme. 2013. Open domain targeted sentiment. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1643–1654.

World Health Organization. 2001. The world health report 2001 - Mental health: New understanding, new hope. Technical report, Genf, Schweiz.

Minsu Park, Chiyoung Cha, and Meeyoung Cha. 2012. Depressive moods of users portrayed in Twitter. In *Proceedings of the ACM SIGKDD Workshop on Healthcare Informatics (HI-KDD)*.

Michael J. Paul and Mark Dredze. 2011. You are what you tweet: Analyzing Twitter for public health. In *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM)*.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, and Matthieu Perrot Édouard Duchesnay. 2011. scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, 12:2825–2830.

Frank J. Penedo and Jason R. Dahn. 2005. Exercise and well-being: a review of mental and physical health benefits associated with physical activity. *Current Opinion in Psychiatry*, 18(2):189–193, March.

James W. Pennebaker, Cindy K. Chung, Molly Ireland, Amy Gonzales, and Roger J. Booth. 2007. *The development and psychometric properties of LIWC2007*.

Nairan Ramirez-Esparza, Cindy K. Chung, Ewa Kacewicz, and James W. Pennebaker. 2008. The psychology of word use in depression forums in English and in Spanish: Testing two text analytic approaches. In *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM)*.

Philip Resnik, Anderson Garron, and Rebecca Resnik. 2013. Using topic modeling to improve prediction of neuroticism and depression. In *Proceedings of the 2013 Conference on Empirical Methods in Natural*, pages 1348–1353.

Stephanie S. Rude, Eva-Maria Gortner, and James W. Pennebaker. 2004. Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, 18(8):1121–1133, December.

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. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLOS One*, 8(9).

Yla R. Tausczik and James W. Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54.

John W. Tukey. 1977. Box-and-whisker plots. *Exploratory Data Analysis*, pages 39–43.

Myrna M. Weissman, Roger C. Bland, Glorisa J. Canino, Carlo Faravelli, Steven Greenwald, Hai-Gwo Hwu, Peter R. Joyce, Eile G. Karam, Chung-Kyoon Lee, Joseph Lellouch, Jean-Pierre Lépine, Stephen C. Newman, Maritza Rubio-Stipec, J. Elisabeth Wells, Priya J. Wickramaratne, Hans-Ulrich Wittchen, and Eng-Kung Yeh. 1996. Cross-national epidemiology of major depression and bipolar disorder. *Journal of the American Medical Association (JAMA)*, 276(4):293–299.

Hadley Wickham. 2009. *ggplot2: elegant graphics for data analysis*. Springer.