From ADHD to SAD: Analyzing the Language of Mental Health on Twitter through Self-Reported Diagnoses

Glen Coppersmith
Qntfy
glen@qntfy.io

Mark Dredze, Craig Harman
Human Language Technology
Center of Excellence
Johns Hopkins University
mdredze|charman@jhu.edu

Kristy Hollingshead
IHMC
kseitz@ihmc.us

Abstract

Many significant challenges exist for the mental health field, but one in particular is a lack of data available to guide research. Language provides a natural lens for studying mental health – much existing work and therapy have strong linguistic components, so the creation of a large, varied, language-centric dataset could provide significant grist for the field of mental health research. We examine a broad range of mental health conditions in Twitter data by identifying self-reported statements of diagnosis. We systematically explore language differences between ten conditions with respect to the general population, and to each other. Our aim is to provide guidance and a roadmap for where deeper exploration is likely to be fruitful.

1 Introduction

A recent study commissioned by the World Economic Forum projected that mental disorders will be the single largest health cost, with global costs increasing to $6 trillion annually by 2030 (Bloom et al., 2011). Since mental health impacts the risk for chronic, non-communicable diseases, in a sense there is "no health without mental health" (Prince et al., 2007). The importance of mental health has driven the search for new and innovative methods for obtaining reliable information and evidence about mental disorders. The WHO's Mental Health Action Plan for the next two decades calls for the strengthening of "information systems, evidence and research," which necessitates new development and improvements in global mental health surveillance capabilities (World Health Organization, 2013).

As a result, research on mental health has turned to web data sources (Ayers et al., 2013; Althouse et al., 2014; Yang et al., 2010; Hausner et al., 2008), with a particular focus on social media (De Choudhury, 2014; Schwartz et al., 2013a; De Choudhury et al., 2011). While many users discuss physical health conditions such as cancer or the flu (Paul and Dredze, 2011; Dredze, 2012; Aramaki et al., 2011; Hawn, 2009), some also discuss mental illness. There are a variety of motivations for users to share this information on social media: to offer or seek support, to fight the stigma of mental illness, or perhaps to offer an explanation for certain behaviors.

Past mental health work has largely focused on depression, with some considering post-traumatic stress disorder (Coppersmith et al., 2014b), suicide (Tong et al., 2014; Jashinsky et al., 2014), seasonal affective disorder, and bipolar disorder (Coppersmith et al., 2014a). While these represent some of the most common mental disorders, it only begins to consider the range of mental health conditions for which social media could be utilized. Yet obtaining data for many conditions can be difficult, as previous techniques required the identification of affected individuals using traditional screening methods (De Choudhury, 2013; Schwartz et al., 2013b).

Coppersmith et al. (2014a) proposed a novel way of obtaining mental health related Twitter data. Using the self-identification technique of Beller et al. (2014), they looked for statements such as "I was diagnosed with depression", automatically uncovering a large number of users with mental health conditions. They demonstrated success at both surveillance and analysis of four mental health conditions. While a promising first step, the technique's efficacy for a larger range of disorders remained untested.

Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, pages 1–10,
Denver, Colorado, June 5, 2015. ©2015 Association for Computational Linguistics

In this paper we employ the techniques of Coppersmith et al. (2014a) to amass a large, diverse collection of social media and associated labels of diagnosed mental health conditions. We consider the broadest range of conditions to date, many significantly less prevalent than the disorders examined previously. This tests the capacity of our approach to scale to many mental health conditions, as well as its capability to analyze relationships between conditions. In total, we present results for ten conditions, including the four considered by Coppersmith et al. (2014a). To demonstrate the presence of quantifiable signals for each condition, we build machine learning classifiers capable of separating users with each condition from control users.

Furthermore, we extend previous analysis by considering approximate age- and gender-matched controls, in contrast to the randomly selected controls in most past studies. Dos Reis and Culotta (2015) found demographic controls an important baseline, as they muted the strength of the measured outcomes in social media compared to a random control group. Using demographically-matched controls allows us to clarify the analysis in conditions where age is a factor, e.g., people with PTSD tend to be older than the average user on Twitter.

Using the ten conditions and control groups, we characterize a broad range of differences between the groups. We examine differences in usage patterns of categories from the Linguistic Inquiry Word Count (LIWC), a widely used psychometrically validated tool for psychology-related analysis of language (Pennebaker et al., 2007; Pennebaker et al., 2001). Depression is the only condition for which considerable previous work on social media exists for comparison, and we largely replicate those previous results. Finally, we examine relationships between the language used by people with various conditions — a task for which comparable data has never before been available. By considering multiple conditions, we can measure similarities and differences of language usage between conditions, rather than just between a condition and the general population.

The paper is structured as follows: we begin with a description of how we gathered and curated the data, then present an analysis of the data's coherence and the quantifiable signals we can extract from it, including a broad survey of observed differences in LIWC categories. Finally, we measure language correlations between pairs of conditions. We conclude with a discussion of some possible future directions suggested by this exploratory analysis.

2 Related Work

There is rich literature on the interaction between mental health and language (Tausczik and Pennebaker, 2010; Ramirez-Esparza et al., 2008; Chung and Pennebaker, 2007; Pennebaker et al., 2007; Rude et al., 2004; Pennebaker et al., 2001). Social media's emergence has renewed interest in this topic, though gathering data has been difficult. Deriving measurable signals relevant to mental health via statistical approaches requires large quantities of data that pair a person's mental health status (e.g., diagnosed with PTSD) to their social media feed.

Successful approaches towards obtaining these data have relied on three approaches: **(1) Crowdsourced surveys**: Some mental health conditions have self-assessment questionnaires amenable to administration over the Internet. Combining this with crowdsource platforms like Amazon's Mechanical Turk or Crowdflower, a researcher can administer relevant mental health questionnaires and solicit the user's public social media data for analysis. This technique has been effectively used to examine depression (De Choudhury, 2013; De Choudhury et al., 2013c; De Choudhury et al., 2013b). **(2) Facebook**: Researchers created an application for Facebook users that administered various personality tests, and as part of the terms of service of the application, granted the researchers access to a user's public status updates. This corpus has been used in a wide range of questions from personality (Schwartz et al., 2013b; Park et al., In press), heart disease (Eichstaedt et al., 2015), depression (Schwartz et al., 2014), and psychological well-being (Schwartz et al., 2013a). **(3) Self-Stated Diagnoses**: Some social media users discuss their mental health publicly and openly, which allows researchers to create rich corpora of social media data from users who have a wide range of mental health conditions. This has been used previously to examine depression, PTSD, bipolar, and seasonal affective disorder (Coppersmith et al., 2014a; Coppersmith et al., 2014b; Hohman et al., 2014). A similar approach has been used to identify new mothers for studying the impact of major life events (De Choudhury et al., 2013a). **(4) Affiliation**: Some rely on a user's affiliation to indicate a mental health condition, such as using posts from a depression forum as a sample of depression (Nguyen et al., 2014).

Other work on mental health and related topics have studied questions that do not rely on an explicit diagnosis, such as measuring the moods of Twitter users (De Choudhury et al., 2011) to measure their affective states (De Choudhury et al., 2012). Outside of social media, research has demonstrated how web search queries can measure population level mental health trends (Yang et al., 2010; Ayers et al., 2013; Althouse et al., 2014).

3 Data

We follow the Twitter data acquisition and curation process of Coppersmith et al. (2014a). This data collection method has been previously validated through replication of previous findings and showing predictive power for real-world phenomena (Coppersmith et al., 2014a; Coppersmith et al., 2014b; Hohman et al., 2014), though there likely is some 'selection bias' by virtue of the fact that the data is collected from social media – specifically Twitter – which may be more commonly used by a subset of the population. We summarize the main points of the data collection method here[1].

We obtain messages with self-reported diagnoses using the Twitter API. Self-reported diagnoses are tweets containing statements like "I have been diagnosed with CONDITION", where CONDITION is one of ten selected conditions (each of which has at least 100 users): Attention Deficit Hyperactivity Disorder (ADHD), Generalized Anxiety Disorder (Anx), Bipolar Disorder, Borderline Personality Disorder (Border), Depression (Dep), Eating Disorders (Eating; includes anorexia, bulimia, and eating disorders not otherwise specified [EDNOS]), obsessive compulsive disorder (OCD), post-traumatic stress disorder (PTSD), schizophrenia (Schizo; to include schizophrenia, schizotypal, schizophreniform) and seasonal affective disorder (Seasonal). We use the common names for these disorders, rather than adhering to a more formal one (e.g., DSM-IV or DSM-5), for two reasons: **(1)** to remain agnostic to the current discussion in clinical psychology around the standards of diagnosis; and **(2)** our classification is based on user statements. While sometimes an obvious mapping exists for user statements to more formal definitions (e.g., "shell shock" equates to today's "PTSD"), other times it is less obvious (e.g., "Anxiety" might refer to generalized anxiety disorder or social anxiety disorder).

Condition	**Users**	**Median**	**Total**
ADHD	102	3273	384k
Anxiety	216	3619	1591k
Bipolar	188	3383	720k
Borderline	101	3330	321k
Depression	393	3306	546k
Eating	238	3229	724k
OCD	100	3331	314k
PTSD	403	3241	1251k
Schizophrenia	172	3236	493k
Seasonal Affective	100	3229	340k

Table 1: The number of users with a genuine statement of diagnosis (verified by a human annotator), their median number of tweets, and total tweets for each condition.

Each self-reported diagnosis was examined by one of the authors to verify that it was a genuine statement of a diagnosis, i.e., excluding jokes, quotes, or disingenuous statements.[2] Previous work shows high inter-annotator agreement ($\kappa = 0.77$) for assessing genuine statements of diagnosis (Coppersmith et al., 2014a). For each author of a genuine diagnosis tweet we obtain a set of their **public** Twitter posts using the Twitter API (at least 100 posts per user, but usually more); we do not have access to private messages. All collected data was publicly posted to Twitter between 2008 and 2015.

3.1 Exclusion and Preprocessing

Our analyses focus on user-authored content; we exclude retweets and tweets with a URL since these often quote text from the link. The text is lowercased and all non-standard characters (e.g., emoji) are converted to a systematic ASCII representation

[1] All uses of these data as reported in this paper have been approved by the relevant Institutional Review Board (IRB).

[2] We did not formally analyze the disingenuous statements, but anecdotally many of the jokes seems to stem from laymens terms and understanding of a condition; for example, "The weather in Maryland is totally bipolar."

via Unidecode[3]. Users were removed if their tweets were not at least 75% English, as determined by the Google Compact Language Detector[4]. To avoid bias, we removed the tweets that were used to manually assess genuine statements of diagnosis. However, other tweets with a self-statement of diagnosis may remain in a user's data. Table 1 summarizes the number of users identified and their median number of tweets for each condition.

3.2 Age- and Gender-Matched Controls

Generally, control groups were formed via random selection of Twitter users. Yet physical and mental health conditions have different prevalence rates depending on age and gender. Dos Reis and Culotta (2015) demonstrated that failing to account for these can yield biased control groups that skew results, so we aim to form approximate age- and gender-matched control groups.

There is a rich literature investigating the influence of age and gender on language (Pennebaker, 2011). Since Twitter does not provide demographic information for users, these insights have been broadly applied to inferring demographic information from social media (Volkova et al., 2015; Fink et al., 2012; Burger et al., 2011; Rao et al., 2011; Rao et al., 2010). We use these techniques to estimate the age and gender of each user so as to select an age- and gender-matched control group. For each user in our mental health collection we obtain age and gender estimates from the tools provided by the World Well-Being Project (Sap et al., 2014)[5]. These tools use lexica derived from Facebook data to identify demographics, and have been shown successful on Twitter data. The tools provide continuous valued estimates for age and gender, so we threshold the gender values to obtain a binary label, and use the age score as is.

We draw our community controls from all the Twitter users who tweeted during a two week period in early 2014 as part of Twitter's 1% 'spritzer' stream. Each user who tweeted in English and whose tweets were public had an equal probability of being included in our pool of controls. From this pool, we identify the closest matching control user in terms of age and gender for each user in the mental health collection. We select controls without replacement so a control user can only be included once. In practice, differences between estimated age of paired users were miniscule.

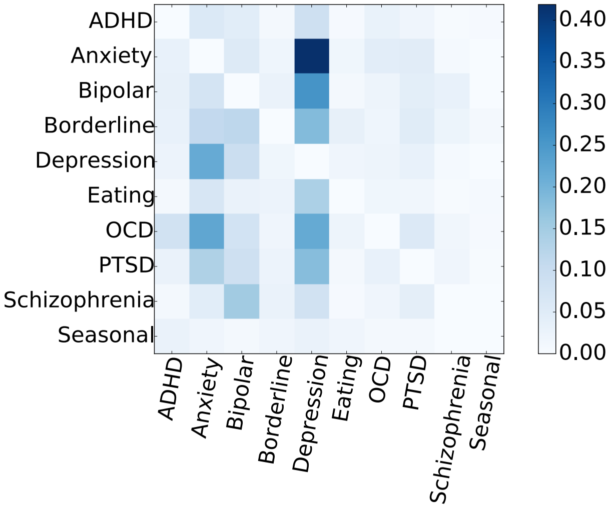

Figure 1: Concomitances or comorbidities: cell color indicates the probability that a user diagnosed with one condition (row) has a concomitant diagnosis of another condition (column). For example: ∼30% of users with schizophrenia also had a diagnosis for bipolar.

3.3 Concomitance and Comorbidity

Concomitant diagnoses are somewhat common in clinical psychology; our data is no different. In cases where a user states a diagnosis for more than one condition, we include them in each condition. For most pairs of conditions, these overlaps are only a small proportion of the data, with a few noted exceptions (e.g., up to 40% of users who have anxiety also have depression, 30% for schizophrenia and bipolar). Figure 1 summarizes the concomitance in our data.

4 Methods and Results

4.1 LIWC differences

We provide a comprehensive picture of differences in usage patterns of LIWC categories between users with various mental health conditions. We measure the proportion of word tokens for each user that falls into a given LIWC category, aggregate by condition, and compare across conditions.

For each user, we calculate the proportion of their tokens that were part of each LIWC category. Thus

[3] https://pypi.python.org/pypi/Unidecode
[4] https://code.google.com/p/cld2/
[5] http://wwbp.org/data.html

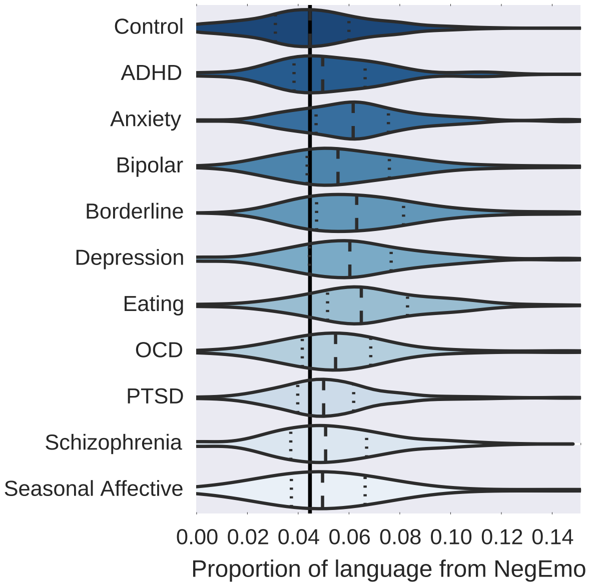

Figure 2: Violin plot showing the frequency of negative emotion LIWC category words by condition. The center dashed line is the median, the dotted line is the inter-quartile-range, and the envelope is an estimate of the distribution. The vertical line is the control group's median.

for each category and each condition, we have an empirical distribution of the proportion of language attributable to that category. The violin plots in Figure 2 show an example of how this changes across conditions as compared to controls.

Table 2 shows deviations for all categories and conditions as follows: '+++' indicates that condition users evince this category significantly more frequently[6] than control users; '+' indicates that the distribution is noticeably higher for the condition population than the control population, but not outside the inter-quartile-range; '−' indicates differences where condition users use this category less frequently than control users.

Some interesting trends emerge from this analysis. First, some categories show differences across a broad range of mental health conditions (e.g., the ANXIETY, AUXILIARY VERBS, COGNITIVE MECHANISMS, DEATH, FUNCTION, HEALTH, and TENTATIVE categories of words). This suggests that there are a subset of changes in language that may be indicative of an underlying mental health condition (without much regard for specificity), while others seem to be very specific to the conditions they are associated with (e.g., INGEST and NEGATIONS with eating disorders). Some of the connections between LIWC categories and mental health conditions have already been substantiated in the mental health literature, while others (e.g., AUXVERB) have not and are ripe for further exploration. Second, many of the conditions show similar patterns (e.g., anxiety, bipolar, borderline, and depression), while others have distinct patterns (e.g., eating disorders and seasonal affective disorder). It is worth emphasizing that a direct mapping between these and previously-reported LIWC results (in, e.g., Coppersmith et al. (2014a) and De Choudhury et al. (2013c)) is not straightforward, since previous work did not use demographically-matched control users.

4.2 Open-vocabulary Approach

Validated and accepted lexicons like LIWC cover a mere fraction of the total language usage on social media. Thus, we also use an open-vocabulary approach, which has greater coverage than LIWC, and has been shown to find quantifiable signals relevant to mental health in the past (Coppersmith et al., 2014a; Coppersmith et al., 2014b). Though many open-vocabulary approaches exist, we opt for one that provides a reasonable score even for very short text, and is robust to the creative spellings, lack of spaces, and other textual *faux pas* common on Twitter: character n-gram language models (CLMs).

In essence, rather than examining words or sequences of words, CLMs examine sequences of characters, including spaces, punctuation, and emoticons. Given a set of data from two classes (in our case, one from a given mental health condition, the other from its matched controls), the model is trained to recognize which sequences of characters are likely to be generated by either class. When these models are presented with novel text, they estimate which of the classes was more likely to have generated it. For brevity we will omit discussion of the exact score calculation and refer the interested reader to Coppersmith et al. (2014a). For all we do here, higher scores will indicate a tweet is more likely to come from a user with a given mental health condition, and lower scores are more likely to come from a control user. Since we are examining ten conditions, we have ten pairs of CLMs (for each pair,

[6] Specifically, the median of the condition distribution is outside the inter-quartile-range of the control distribution.

LIWC	ADHD	Anx	Bipolar	Border	Dep	Eating	OCD	PTSD	Schizo	Seasonal
Funct	+++	+++	+++	+++	+++	+++	+++	+	+++	+++
Pronoun		+			+	+++	+			
PPron		+					+			
I		+			+	+++	+++			
We		-	-	—	-	—				
They	+++		+	+				+	+	
IPron	+++	+		+			+++			
Article						-		+	+++	+
Verb				+	+	+++	+			
AuxVerb	+	+++	+++	+++	+++	+++	+++	+	+++	+
Past			+							+
Present					+	+++				
Adverb		+				+++	+			
Conj*	+	+++	+	+++	+++	+++	+++		+	+++
Negate						+				
Quant	+	+		+++			+	+	+++	
Swear				+		+	+			
PosEmo							-	-		-
NegEmo		+++	+	+++	+	+++				
Anxiety	+	+++	+	+++	+	+++	+++	+	+	+
Anger		+	+	+++	+	+++	+			
Sad				+		+++	+			
CogMech	+++	+++	+++	+++	+++	+++	+++	+	+++	
Insight	+++	+				+++	+++	+	+	
Cause	+++	+	+	+	+	+++	+++	+	+	
Discrep		+				+++				
Tentat	+++	+++	+	+++	+++		+++	+	+++	
Incl				+						+++
Excl	+++	+++	+	+++	+++	+++	+++			
Feel						+				
Bio		+	+	+		+++	+			
Body						+				
Health	+	+++	+++	+++	+	+++	+++	+	+	
Ingest						+				
Relativ	—						-	-	-	
Motion	-	-	-	—	-	-	—	—	—	
Space						-				+
Time	-			-				+++	+++	
Leisure				-	-	—		-	-	
Home				-					-	
Death	+	+++	+	+++	+	+	+	+	+++	
Assent	-								-	
Pro1		+			+	+++	+++			
Pro3	+							+		
LIWC	**ADHD**	**Anx**	**Bipolar**	**Border**	**Dep**	**Eating**	**OCD**	**PTSD**	**Schizo**	**Seasonal**

Table 2: Full list of deviations by LIWC category for each condition. Category names that are *'d may have been affected by our normalization and tokenization procedure. Categories for which no significant differences were observed: Achieve, Affect, Certain, Family, Filler*, Friend, Future, Hear, Humans, Inhibition, Money, Nonfluencies, Number, Perceptual, Prepositions, Pro2, Religion, See, Sexual, SheHe, Social.

one CLM is trained from the users with a given mental health condition, and one CLM is trained from their matched controls).

4.3 Quantifiable Differences

To validate that our CLMs are capturing quantifiable differences relevant to their specific conditions, we examine their accuracy on a heldout set of users. Each condition-specific CLM produces a score that roughly equates to how much more (or less) likely it is to have come from a user with the given condition (e.g., PTSD) than a control. We aggregate these scores to compute a final score for use in classification. We score each tweet with the CLM and use the score to make a binary distinction – is this tweet more likely to have been generated by someone who has PTSD or a control? We calculate the proportion of these tweets that are classified as PTSD-like (the *overall mean*), which can be thought of as how PTSD-like this user looks over all time. Given that some of these symptoms change with time, we can also compute a more localized version of this mean, and derive a score according to the "most PTSD-like period the user has". This is done by ordering these binary decisions by the time the tweet was authored, selecting a window of 50 tweets, and calculating the proportion of those tweets classified as PTSD-like. We then slide this window one tweet further (removing the oldest tweet, and adding in the next in the user's timeline) and calculate the proportion again. The highest this rolling-window mean achieves will be referred to as the *maximum local mean*. We combine these scores to yield the classifier score $\psi = \textit{overall mean} * \textit{maximum local mean}$, capturing how PTSD-like the user is over all time, and how PTSD-like they are at their most severe.

We estimated the performance of our classifiers for each condition on distinguishing users with a mental health condition from their community controls via 10-fold cross-validation. This differs only slightly from standard cross-fold validation in that our observations are paired; we maintain this pairing when assigning folds – each mental health condition user and their matched control are in the same fold. To assess performance, we could draw a line (a threshold) in the ranked list, and classify all users above that line as having the mental health condition, and all users below that line as controls. Those with the condition above the line would be correctly classified (hits), while those controls above the line would be incorrectly classified (false alarms). Figure 3 shows performance of this classifier as Receiver Operating Characteristic (ROC) curves as we adjust this threshold, one curve per mental health condition. The x-axis shows the proportion of false alarms and the y-axis shows the proportion of true hits. All our classifiers are better than chance, but far from perfect. To aid interpretation, Table 3 shows precision at 10% false alarms.

Performance for most conditions is reasonable, except seasonal affective disorder which is very difficult (as was reported by Coppersmith et al. (2014a)). Anxiety and eating disorders have much better performance than the other conditions. Most

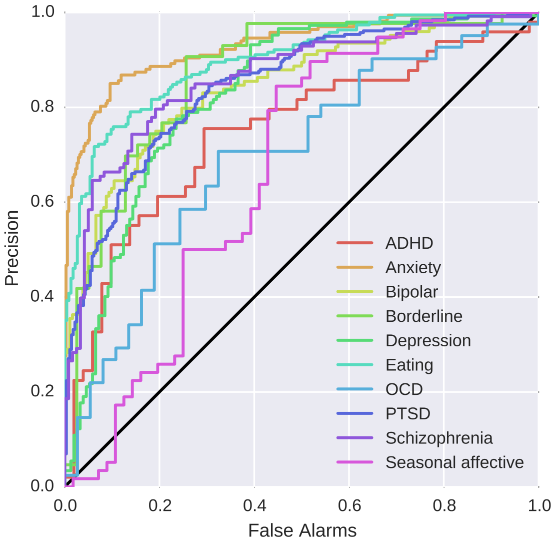

Figure 3: ROC curves for distinguishing diagnosed from control users, for each of the disorders examined. Chance performance is indicated by the black diagonal line.

Condition	Precision
ADHD	52%
Anxiety	85%
Bipolar	63%
Borderline	58%
Depression	48%
Eating	76%
OCD	27%
PTSD	55%
Schizophrenia	67%
Seasonal Affective	5 %

Table 3: Classifier precision with 10% false alarms.

importantly, though, for all conditions (including seasonal affective disorder), we are able to identify language usage differences from control groups.

4.4 Cross Condition Comparisons

Given the breadth of our language data, we can compare across mental health conditions, examining relationships between the conditions under investigation, rather than only how each condition differs from controls. Previous work (Coppersmith et al., 2014a) reported preliminary findings that indicated a possible relationship between the language use from different mental health conditions: similar conditions (either in concomitance and comorbidity or symptomatology) had similar language. The story found here is related, but more complicated. For this comparison, we build new CLMs that *exclude* any user with a concomitant disorder (to prevent their data from making their conditions appear artificially similar). We then score a random sample of 1 million tweets that meet our earlier filters with the CLMs from each condition. We could then examine how the language in any pair of conditions is related by calculating the Pearson's correlation (r) between the scores from these models.

More interesting, though, is how all these conditions relate to one another, rather than any given pair. To that end, we use a standard clustering algorithm[7], shown in Figure 4. Here, each condition is represented by a vector of its Pearson's r correlations, calculated as above, to each of the conditions (to include an $r = 1.0$ to itself). Each condition starts as its own cluster on the left side of the figure. Moving to the right, clusters are merged, most similar first, until all conditions merge into a single cluster. One particular clustering is highlighted by the colors: conditions with blue lines are in clusters of their own, so seasonal affective, ADHD, and borderline appear to be significantly different from the rest); and schizophrenia and OCD are clustered together, shown in red. While this is not the most obvious grouping of conditions, the patterns are far from random: the disorders in green (PTSD, bipolar, eating disorders, anxiety, and depression) have somewhat frequent concomitance in our data and elsewhere (Kessler et al., 2005) and recent research indicates links between OCD and schizophrenia (Meier et al., 2014). Notably, these data are not age- and gender-matched, so these variables also likely factor into the clustering. Thus, we leave this particular relationship between language and mental health as an open question, suggesting fertile grounds for more controlled future work.

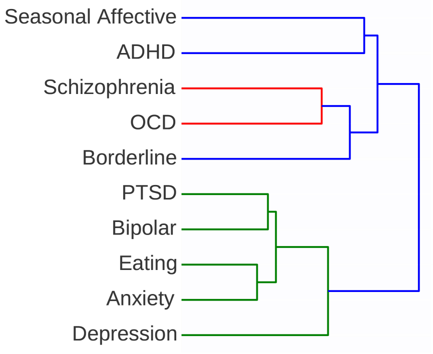

Figure 4: Hierarchical clustering dendrogram of conditions clustered according to the similarity of their users' language. Distance between merged clusters increases monotonically with the level of the merger; thus lower merges (further to the left) indicate greater similarity (e.g., language usage from Seasonal Affective and ADHD users is very different from conditions in the green cluster, given how far right the red merge-point is).

5 Conclusion

We examined the language of social media from users with a wide range of mental health conditions, providing a roadmap for future work. We explored simple classifiers capable of distinguishing these users from their age- and gender-matched controls, based on signals quantified from the users' language. The classifiers also allowed us to systematically compare the language used by those with the ten conditions investigated, finding some groupings of the conditions found elsewhere in the literature, but not altogether obvious. We take this as evidence that examining mental health through the lens of language is fertile ground for advances in mental health writ large. The wealth of information encoded in continually-generated social media is ripe for analysis – data scientists, computational linguists, and clinical psychologists, together, are well positioned to drive this field forward.

[7] Hierarchical, agglomerative clustering from Python's scipy.hierarchy.linkage (Jones et al., 2001).

Acknowledgments

The authors would like to thank Bradley Skaggs, Matthew DiFabion, and Aleksander Yelskiy for their insights throughout this endeavor.

References

Benjamin M. Althouse, Jon-Patrick Allem, Matthew A. Childers, Mark Dredze, and John W. Ayers. 2014. Population health concerns during the United States' great recession. *American Journal of Preventive Medicine*, 46(2):166–170.

Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. 2011. Twitter catches the flu: Detecting influenza epidemics using Twitter. In *Proceedings of 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.

Charley Beller, Rebecca Knowles, Craig Harman, Shane Bergsma, Margaret Mitchell, and Benjamin Van Durme. 2014. I'm a Belieber: Social roles via self-identification and conceptual attributes. In *Proceedings of ACL*.

David E. Bloom, Elizabeth Cafiero, Eva Jané-Llopis, Shafika Abrahams-Gessel, Lakshmi Reddy Bloom, Sana Fathima, Andrea B. Feigl, Tom Gaziano, Ali Hamandi, Mona Mowafi, Ankur Pandya, Klaus Prettner, Larry Rosenberg, Ben Seligman, Adam Z. Stein, and Cara Weinstein. 2011. The global economic burden of non-communicable diseases. Technical report, Geneva: World Economic Forum.

John D. Burger, John Henderson, George Kim, and Guido Zarrella. 2011. Discriminating gender on Twitter. In *Proceedings of EMNLP*.

Cindy Chung and James Pennebaker. 2007. The psychological functions of function words. *Social Communication*.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014a. Quantifying mental health signals in Twitter. In *Proceedings of the ACL Workshop on Computational Linguistics and Clinical Psychology*.

Glen Coppersmith, Craig Harman, and Mark Dredze. 2014b. Measuring post traumatic stress disorder in Twitter. In *Proceedings of ICWSM*.

Munmun De Choudhury, Scott Counts, and Michael Gamon. 2011. Not all moods are created equal! Exploring human emotional states in social media. In *Proceedings of ICWSM*.

Munmun De Choudhury, Michael Gamon, and Scott Counts. 2012. Happy, nervous or surprised? Classification of human affective states in social media. In *Proceedings of ICWSM*.

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 16th ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW)*.

Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013b. Social media as a measurement tool of depression in populations. In *Proceedings of the 5th ACM International Conference on Web Science*.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013c. Predicting depression via social media. In *Proceedings of ICWSM*.

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*.

Munmun De Choudhury. 2014. Can social media help us reason about mental health? In *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web (WWW)*.

Virgile Landeiro Dos Reis and Aron Culotta. 2015. Using matched samples to estimate the effects of exercise on mental health from Twitter. In *Proceedings of AAAI*.

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

Johannes C. Eichstaedt, Hansen Andrew Schwartz, Margaret L. Kern, Gregory Park, Darwin R. Labarthe, Raina M. Merchant, Sneha Jha, Megha Agrawal, Lukasz A. Dziurzynski, Maarten Sap, Christopher Weeg, Emily E. Larson, Lyle H. Ungar, and Martin E. P. Seligman. 2015. Psychological language on Twitter predicts county-level heart disease mortality. *Psychological Science*, 26(2):159–169.

Clayton Fink, Jonathon Kopecky, and Maksym Morawski. 2012. Inferring gender from the content of tweets: A region specific example. In *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media (ICWSM)*.

Helmut Hausner, Göran Hajak, and Hermann Spießl. 2008. Gender differences in help-seeking behavior on two internet forums for individuals with self-reported depression. *Gender Medicine*, 5(2):181–185.

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.

Elizabeth Hohman, David Marchette, and Glen Coppersmith. 2014. Mental health, economics, and population in social media. In *Proceedings of JSM*.

Jared Jashinsky, Scott H Burton, Carl L Hanson, Josh West, Christophe Giraud-Carrier, Michael D Barnes, and Trenton Argyle. 2014. Tracking suicide risk factors through Twitter in the U.S. *Crisis*, 35(1).

Eric Jones, Travis Oliphant, Pearu Peterson, et al. 2001. SciPy: Open source scientific tools for Python. [Online; accessed 2015-03-11].

R. C. Kessler, W. T. Chiu, O. Demler, and E. E. Walters. 2005. Prevalence, severity, and comorbidity of twelve-month DSM-IV disorders in the National Comorbidity Survey Replication (NCS-R). *Archives of General Psychiatry*, 62(6):617–627.

Sandra M. Meier, Liselotte Petersen, Marianne G. Pedersen, Mikkel C.B. Arendt, Philip R. Nielsen, Manuel Mattheisen, Ole Mors, and Preben B. Mortensen. 2014. Obsessive-compulsive disorder as a risk factor for schizophrenia: a nationwide study. *JAMA psychiatry*, 71(11):1215–1221.

Thin Nguyen, Dinh Phung, Bo Dao, Svetha Venkatesh, and Michael Berk. 2014. Affective and content analysis of online depression communities. *IEEE Transactions on Affective Computing*, 5(3):217–226.

Greg Park, H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, David J. Stillwell, Michal Kosinski, Lyle H. Ungar, and Martin E. P. Seligman. In press. Automatic personality assessment through social media language. *Journal of Personality and Social Psychology*.

Michael J. Paul and Mark Dredze. 2011. You are what you tweet: Analyzing Twitter for public health. In *Proceedings of ICWSM*.

James W. Pennebaker, Martha E. Francis, and Roger J. Booth. 2001. *Linguistic inquiry and word count: LIWC2001*. Erlbaum Publishers, Mahwah, NJ.

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

James W. Pennebaker. 2011. The secret life of pronouns. *New Scientist*, 211(2828):42–45.

Martin Prince, Vikram Patel, Shekhar Saxena, Mario Maj, Joanna Maselko, Michael R. Phillips, and Atif Rahman. 2007. No health without mental health. *The Lancet*, 370(9590):859–877.

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 ICWSM*.

Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. 2010. Classifying latent user attributes in Twitter. In *Proceedings of the 2nd International Workshop on Search and Mining User-generated Contents*.

Delip Rao, Michael J. Paul, Clayton Fink, David Yarowsky, Timothy Oates, and Glen Coppersmith. 2011. Hierarchical bayesian models for latent attribute detection in social media. In *Proceedings of ICWSM*.

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.

Maarten Sap, Greg Park, Johannes C. Eichstaedt, Margaret L. Kern, David J. Stillwell, Michal Kosinski, Lyle H. Ungar, and H. Andrew Schwartz. 2014. Developing age and gender predictive lexica over social media. In *Proceedings of EMNLP*, pages 1146–1151.

H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Richard E. Lucas, Megha Agrawal, Gregory J. Park, Shrinidhi K. Lakshmikanth, Sneha Jha, Martin E. P. Seligman, and Lyle H. Ungar. 2013a. Characterizing geographic variation in well-being using tweets. In *Proceedings of ICWSM*.

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

H. Andrew Schwartz, Johannes Eichstaedt, Margaret L. Kern, Gregory Park, Maarten Sap, David Stillwell, Michal Kosinski, and Lyle Ungar. 2014. Towards assessing changes in degree of depression through facebook. In *Proceedings of the ACL Workshop on Computational Linguistics and Clinical Psychology*.

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.

Christopher M Homan Tong, Ravdeep Johar, Liu Cecilia, Megan Lytle, Vincent Silenzio, and Cecilia O. Alm. 2014. Toward macro-insights for suicide prevention: Analyzing fine-grained distress at scale. In *Proceedings of the ACL Workshop on Computational Linguistics and Clinical Psychology*.

Svitlana Volkova, Yoram Bachrach, Michael Armstrong, and Vijay Sharma. 2015. Inferring latent user properties from texts published in social media. In *Proceedings of AAAI*.

World Health Organization. 2013. *Mental health action plan 2013-2020*. Geneva: World Health Organization.

Albert C. Yang, Norden E. Huang, Chung-Kang Peng, and Shih-Jen Tsai. 2010. Do seasons have an influence on the incidence of depression? The use of an internet search engine query data as a proxy of human affect. *PLOS ONE*, 5(10):e13728.