Measuring Post Traumatic Stress Disorder in Twitter

Glen A. Coppersmith Craig T. Harman Mark H. Dredze
Human Language Technology Center of Excellence
Johns Hopkins University
Balitmore, MD, USA

Abstract

Traditional mental health studies rely on information primarily collected through personal contact with a health care professional. Recent work has shown the utility of social media data for studying depression, but there have been limited evaluations of other mental health conditions. We consider post traumatic stress disorder (PTSD), a serious condition that affects millions worldwide, with especially high rates in military veterans. We also present a novel method to obtain a PTSD classifier for social media using simple searches of available Twitter data, a significant reduction in training data cost compared to previous work. We demonstrate its utility by examining differences in language use between PTSD and random individuals, building classifiers to separate these two groups and by detecting elevated rates of PTSD at and around U.S. military bases using our classifiers.

Introduction

Mental health conditions affect a significant percentage of the U.S. adult population each year, including depression (6.7%), eating disorders like anorexia and bulimia (1.6%), bipolar disorder (2.6%) and post traumatic stress disorder (PTSD) (3.5%).[1] PTSD and other mental illnesses are difficult to diagnose, with competing standards for diagnosis based on self-reports and testimony from friends and relatives.[2] In recent years, several studies have turned to social media data to study mental health, since it provides an unbiased collection of a person's language and behavior, which has been shown to be useful in diagnosing conditions (De Choudhury 2013). Additionally, from a public health standpoint, social media data and Web data in general have enabled large scale analyses of a population's health status beyond what has previously been possible with traditional methods (Ayers et al. 2013).

While social media provides ample data for many types of public health analysis (Paul and Dredze 2011), mental health studies still face serious challenges. First, other health work in social media, such as disease surveillance (Brownstein, Freifeld, and Madoff 2009; Chew and Eysenbach 2010; Lamb, Paul, and Dredze 2013) and modeling (Sadilek, Kautz, and Silenzio 2012), rely on explicit mentions of illness or health issues; if people are sick, they say so. In contrast, mental health conditions largely display implicit changes in language and behavior, such as a switch in the types of topics, a shift in word usage or a shift in frequency of posts. While De Choudhury et al. (2013) find some examples of explicit depression mentions, the focus is on more subtle changes in language (e.g., pronoun use).

Second, obtaining labeled data for a mental health condition is challenging since we are examining implicit features of language. De Choudhury et al. (2013) rely on (crowdsourced) volunteers to take depression surveys and offer their Twitter feed for research. While this yields reliable data, it is time-consuming and challenging to build large data sets for a diverse set of mental health conditions. Furthermore, the necessary mental health evaluations such as the DSM (Diagnostic and Statistical Manual of Mental Disorders)[3], are difficult to perform as these evaluations require a trained diagnostician and have been criticized as unscientific and subjective (Insel 2013). Thus, relying on data from crowdsourced volunteers to build datasets of users with diverse mental health conditions is difficult, and perhaps untenable. We provide an alternate method for gathering samples that partially ameliorate these problems – ideally to be used in concert with existing methods.

In this paper, we study PTSD in Twitter data, one of the first studies to consider social media for a mental health condition beyond depression (De Choudhury, Counts, and Horvitz 2013; De Choudhury et al. 2013; Rosenquist, Fowler, and Christakis 2010). Rather than rely on traditional PTSD diagnostic tools (Foa 1995) for finding data, we demonstrate that some PTSD users can be easily and automatically identified by scanning for tweets expressing explicit diagnoses. While it is natural to be suspicious of self-identified reporting, we find that self-identifying PTSD users have demonstrably different language usage patterns from the random users, according to the Linguistic Inquiry Word Count (LIWC), a psychometrically validated analysis tool (Pennebaker et al. 2007). We demonstrate elsewhere (Coppersmith, Dredze, and Harman 2014) that data obtained

[1] www.nimh.nih.gov/health/publications/the-numbers-count-mental-disorders-in-america

[2] en.wikipedia.org/wiki/List_of_diagnostic_classification_and_rating_scales_used_in_psychiatry

[3] en.wikipedia.org/wiki/Diagnostic_and_Statistical_Manual_of_Mental_Disorders

in this way replicates analyses performed via LIWC on the crowdsourced survey respondents of De Choudhury et al. (2013). We also demonstrate that users who self-identify are measurably different from random users by learning a classifier to discriminate between self-identified and random users. We further show how this data can be used to train a classifier that detects elevated incidences of PTSD in tweets from U.S. military bases as compared to the general U.S. population, with a further increase around bases that deployed combat troops overseas. We intend for this initial finding (which is small, but statistically significant) to be a demonstration of the types of analysis Twitter data enables for public health. Given the small effect size, replication and further study are called for.

Data

We used an automated analysis to find potential PTSD users, and then refined the list manually. First, we had access to a large multi-year historical collection from the Twitter keyword streaming API, where keywords were selected to focus on health topics. We used a regular expression[4] to search for statements where the user self-identifies as being diagnosed with PTSD. The 477 matching tweets were manually reviewed to determine if they indicated a genuine statement of a diagnosis for PTSD. Table 1 shows examples from the 260 tweets that indicated a PTSD diagnosis.

Next, we selected the username that authored each of these tweets and retrieved up to the 3200 most recent tweets from that user via the Twitter API. We then filtered out users with less than 25 tweets and those whose tweets were not at least 75% in English (measured using an automated language ID system.) This filtering left us with 244 users as positive examples.

We repeated this process for a group of randomly selected users. We randomly selected 10,000 usernames from a list of users who posted to our historical collection within a selected two week window. We then downloaded all tweets from these users. After filtering (as above) 5728 random users remain, whose tweets were used as negative examples.

Methods

We use our positive and negative PTSD data to train three classifiers: one unigram language model (ULM) examining individual whole words, one character n-gram language model (CLM), and one from the LIWC categories above. The LMs have been shown effective for Twitter classification tasks (Bergsma et al. 2012) and LIWC has been previously used for analysis of mental health in Twitter (De Choudhury et al. 2013). The language models measure the probability that a word (ULM) or a string of characters (CLM) was generated by the same underlying process as the training data. Here, one of each language model (clm^+ and ulm^+) is trained from the tweets of PTSD users, and a second (clm^- and ulm^-) from the tweets from random users. Each test tweet t is scored by comparing proabilities from each LM:

$$s = \frac{lm^+(t)}{lm^-(t)} \quad (1)$$

A threshold of 1 for s divides scores into positive and negative classes. In a multi-class setting, the algorithm minimizes the cross entropy, selecting the model with the highest probability. For each user, we calculate the proportion of tweets scored positively by each LIWC category. These proportions are used as a feature vector in a loglinear regression model (Pedregosa et al. 2011).

Prior to training, we preprocess the text of each tweet: we replaced all usernames with a single token (USER), lowercased all text, and removed extraneous whitespace. We also excluded any tweet that contained a URL, as these often pertain to events external to the user (e.g., national news stories). In total, we used 463k PTSD tweets and sampled 463k non-PTSD tweets to create a balanced data set.

Results

PTSD Language

Numerous studies have investigated the language that PTSD sufferers use in "trauma narratives" describing their traumatic experiences (for a review, see (O'Kearney and Perrott 2006)), but only a handful of these studies have used automated analysis of language. The Linguistic Inquiry Word Count (LIWC) (Pennebaker et al. 2007) has been used to analyze the narratives of 28 female assault victims being treated for chronic PTSD, and found that the LIWC cognitive words category was inversely correlated with post-treatment anxiety, and that social adjustment was negatively related to negative emotion words and death words (Alvarez-Conrad, Zoellner, and Foa 2001).

We conduct a LIWC analysis of the PTSD and non-PTSD tweets to determine if there are differences in the language usage of PTSD users. We applied the LIWC battery and examined the distribution of words in their language. Each tweet was tokenized by separating on whitespace. For each user, for a subset of the LIWC categories, we measured the proportion of tweets that contained at least one word from that category. Specifically, we examined the following nine categories: first, second and third person pronouns, swear, anger, positive emotion, negative emotion, death, and anxiety words. Second person pronouns were used significantly less often by PTSD users, while third person pronouns and words about anxiety were used significantly more often. Unsurprisingly, given the stylistic and topical difference between tweets and trauma narratives, we do not observe the same trends as Alvarez-Conrad et al. (2001), but the fact that significant differences in language use exist is an encouraging demonstration of the effectiveness of our data.

Classification Accuracy

Previous work on mental health in social media used validated psychological surveys to obtain mental health labels for users (De Choudhury et al. 2013). In comparison, our data labeling method is much faster and easier, though the question remains: are the obtained labels reliable? We answer this question by evaluating various classifiers in terms

[4]Case insensitive regex:
`\Wptsd\W|\Wp\.t\.s\.d\.\W|post[- ]traumatic[- ]stress[- ]disorder[- ]`

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
@USERNAME 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

Table 1: Examples of tweets expressing a PTSD diagnosis.

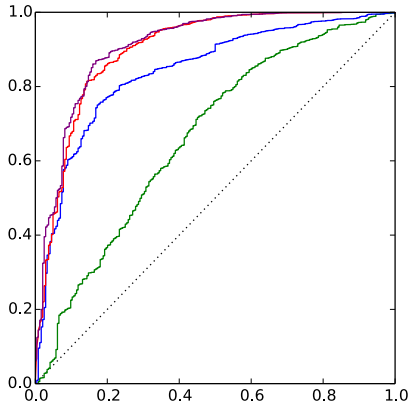

Figure 1: ROC curves for classifying PTSD and non-PTSD users: y-axis is the proportion of correct detections and (x-axis) are false alarms. ULM is in red, CLM in blue, ULM+CLM in purple, and LIWC in green. Chance performance is the black dotted line.

of their ability to differentiate PTSD and random users. If a classifier can learn to differentiate these users, then we can infer that it is finding a useful signal from the data. If the labels are unreliable, we would expect random performance from the classifier.

We evaluated the classifiers via leave-one-out cross validation setting in both a balanced and a non-balanced dataset. In the balanced data set, a single PTSD and non-PTSD user is left out. In the non-balanced setting, each fold held out a single PTSD user and non-PTSD users proportional to the overall ratio between positive and negative training examples, ensuring identical ratios in each training fold. Leave-one-out cross validation provides maximum training data while evaluating every user in turn. We obtained different operating points by varying the classification threshold for s. The results are shown in Figure 1, a receiver operating characteristic (ROC) curve for the proportion of correct detection (y-axis) against the proportion of false alarms (x-axis). In decreasing order of performance is ULM, CLM, and finally LIWC. The non-random performance of our classifiers at separating these classes is further evidence that the data collection method yields sensible data. This additionally indicates that there is more linguistic signal relevant to the separation of users than is captured by LIWC alone.

PTSD in the Military

Previous work on detecting health trends for Twitter have shown that imperfect classifiers can still inform us about public health trends. We consider whether our classifier can detect different incidence levels of PTSD in populations with different risk factors, where a higher incidence of PTSD

Frequently Deployed	Less Frequently Deployed	Urban	Suburban/Rural
Fort Benning (GA)	Arnold Air Force Base (TN)	Boston (MA)	Amherst (MA)
Fort Bragg (NC)	Fort Leavenworth (KS)	Cincinnati (OH)	Cape Cod (MA)
Fort Campbell(KY)	Fort Sill (OK)	Detroit (MI)	Fingerlakes Region (NY)
Fort Carson (CO)	McGuire-Dix-Lakehurst (NJ)	Milwaukee (WI)	Laramie (WY)
Fort Drum (NY)	MacDill Air Force Base (FL)	Minneapolis (MN)	Moab (UT)
Fort Hood (TX)	Maxwell Air Force Base (AL)	Pittsburgh (PA)	Newry (ME)
Fort Irwin (CA)	Offutt Air Force Base (NE)	Portland (OR)	Panacea (FL)
Fort Lewis (WA)	Scott Air Force Base (IL)	Seattle (WA)	Quechee Gorge (VT)
Fort Riley (KS)	Wright-Patterson AFB (OH)	St. Louis (MO)	Red River Gorge (KY)

Table 2: The locations used for the military experiment.

tweets could indicate both an elevated awareness of PTSD, as well as tweets from users with the condition.

Since U.S. military personnel have a higher PTSD rate than the general population, we compared these two populations. However, we do not have a mechanism for separating Twitter accounts of military personnel from the general population, so we instead focused on a geographic division. We selected geographic regions in the U.S. that housed troops recently involved as 'boots on the ground' in the conflicts overseas. Since statistics for deployments are not widely available, we asked a retired service member (not among the authors) who served overseas during the recent conflicts to select installations that would represent this sample, as well as military installations that deployed less during recent conflicts or were not used as 'boots on the ground' since we would expect a lower, but still elevated, rate of PTSD. For each base, we created a bounding box that covered the full base and as little of the surrounding area as possible. These bounding boxes resulted in an average of 407k tweets per box (median 358k, standard deviation 296k) in 2013.

To represent the civilian population, we selected both urban and rural areas across the spectrum of civilian life: major cities, vacation spots, and rural townships (Table 2). These bounding boxes resulted in an average of 711k tweets per box (median 494k, standard deviation 774k) in 2013.

We collected all geocoded tweets in our bounding boxes during 2013 (Twitter location streaming API). We used our CLM classifier to identify PTSD tweets in these bounding boxes, and computed the incidence rate by normalizing by the total number of tweets in the bounding box.

We compare the cartesian product of military bases and civilian areas, noting for each comparison which incidence is greater. Our null hypothesis is that the incidence of PTSD-like tweets is equivalent across military and civilian areas. In 248 out of 342 comparisons military areas have higher PTSD incidence than civilian areas. A binomial test indicates this is statistically significant ($p < 3 \times 10^{-17}$), rejecting the null hypothesis. Our secondary hypothesis – frequently-deploying installations have higher PTSD rates than less-frequently-deploying locations – is tested similarly In 53 of 81 comparisons, the frequently-deploying installations have higher incidence rates ($p = 0.007$). Finally, we

expect no difference between rural and urban areas; indeed, the rates were not statistically significant ($p > 0.8$). This sort of analysis is difficult using traditional population-analysis methods, so predictions from the literature are few, except work comparing PTSD veterans in rural and urban areas, which found no significant differences (Elhai et al. 2004).

Though statistically significant, this effect is small with approximately 1% more PTSD-like tweets in military areas than civilian areas and approximately 0.7% more PTSD-like tweets in frequently-deploying military areas as compared to less-frequently deploying areas. Thus, this result requires further study and replication, but it is suggestive of exciting new avenues for population level mental health research.

Conclusion

Mental health is a growing problem, and one for which social media plays a unique role. We have presented the first analysis of social media for the study of individuals with post traumatic stress disorder. Unlike most previous work, our labeled dataset comes from automated searching of raw Twitter data followed by manual curation. Using a classification task, we demonstrate that this dataset captures real differences between PTSD and non-PTSD users. Furthermore, we analyzed our data using the standard LIWC battery and found statistically significant differences in language use. Finally, we used one of our PTSD classifiers to identify and evaluate trends of PTSD incidence in and around U.S. military installations, with an even higher rate in populations more likely to have been deployed into combat – a statistically significant finding, but with a small effect size. In light of this, We treat this finding cautiously – it should be replicated with other classifiers, more geographic regions, and/or a more explicit group of military users but it is at least suggestive of the sort of population-level analysis enabled by this data and these techniques.

There remain several important open questions. Do users who self-report diagnoses differ from other diagnosed individuals, perhaps sharing more relevant mental health information? What other mental health conditions can be studied using our approach of identifying self-diagnoses? Finally, what opportunities exist for interventions with identified users? What linguistic signals are present in social media but not captured by LIWC? Pursuing the answers to these questions will provide many exciting opportunities for mental health research and help to address this serious public health concern.

Acknowledgments

We would like to thank Kristy Hollingshead for her insightful comments and thoughtful contributions. We would also like to thank Matthew DiFabion for the insight he provided to this work from his time spent honorably serving in the United States Air Force.

References

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

Ayers, J. W.; Althouse, B. M.; Allem, J.-P.; Rosenquist, J. N.; and Ford, D. E. 2013. Seasonality in seeking mental health information on Google. *American Journal of Preventive Medicine* 44(5):520–525.

Bergsma, S.; McNamee, P.; Bagdouri, M.; Fink, C.; and Wilson, T. 2012. Language identification for creating language-specific Twitter collections. In *ACL Workshop on Language in Social Media*.

Brownstein, J.; Freifeld, C.; and Madoff, L. 2009. Digital disease detection — harnessing the web for public health surveillance. *New England Journal of Medicine* 360:2153–2157.

Chew, C., and Eysenbach, G. 2010. Pandemics in the age of Twitter: content analysis of tweets during the 2009 H1N1 outbreak. *PloS one* 5(11):e14118.

Coppersmith, G.; Dredze, M.; and Harman, C. 2014. Quantifable mental health signals in Twitter. Submitted.

De Choudhury, M.; Gamon, M.; Counts, S.; and Horvitz, E. 2013. Predicting depression via social media. In *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media (ICWSM)*.

De Choudhury, M.; Counts, S.; and Horvitz, E. 2013. Predicting postpartum changes in emotion and behavior via social media. In *CHI*, 3267–3276. ACM.

De Choudhury, M. 2013. Role of social media in tackling challenges in mental health. In *Proceedings of the 2nd international workshop on Socially-aware multimedia*, SAM '13, 49–52. New York, NY, USA: ACM.

Elhai, J. D.; Baugher, S. N.; Quevillon, R. P.; Sauvageot, J.; and Frueh, B. C. 2004. Psychiatric symptoms and health service utilization in rural and urban combat veterans with posttraumatic stress disorder. *The Journal of nervous and mental disease* 192(10):701–704.

Foa, E. B. 1995. Post-traumatic Stress Diagnostic Scale (PDS). *Minneapolis: National Computer Systems*.

Insel, T. 2013. Transforming diagnosis. `http://www.nimh.nih.gov/about/director/2013/transforming-diagnosis.shtml`.

Lamb, A.; Paul, M. J.; and Dredze, M. 2013. Separating fact from fear: Tracking flu infections on Twitter. In *North American Chapter of the Association for Computational Linguistics (NAACL)*.

O'Kearney, R., and Perrott, K. 2006. Trauma narratives in posttraumatic stress disorder: a review. *Journal of traumatic stress* 19(1):81–93.

Paul, M. J., and Dredze, M. 2011. You are what you tweet: Analyzing Twitter for public health. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM)*.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; and Duchesnay, M. P. É. 2011. scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research* 12:2825–2830.

Pennebaker, J. W.; Chung, C. K.; Ireland, M.; Gonzales, A.; and Booth, R. J. 2007. The development and psychometric properties of LIWC2007. *Austin, TX, LIWC. Net*.

Rosenquist, J. N.; Fowler, J. H.; and Christakis, N. A. 2010. Social network determinants of depression. *Molecular psychiatry* 16(3):273–281.

Sadilek, A.; Kautz, H. A.; and Silenzio, V. 2012. Modeling spread of disease from social interactions. In *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media (ICWSM)*.