Analyzing Machine Learning Models that Predict Mental Illnesses from Social Media Text

Janith Weerasinghe
Drexel University
bnw37@drexel.edu

Kediel Morales
Drexel University
km3556@drexel.edu

Rachel Greenstadt
Drexel University
rachel.a.greenstadt@drexel.edu

1. INTRODUCTION

Previous studies, both in psychology and linguistics, have shown that individuals with mental illnesses show deviations from normal language use, that these differences can be used to make predictions, and used as a diagnostic tool. Recent studies have shown that machine learning can be used to predict people with mental illnesses based on their writing. However, little attention is paid to the interpretability of the machine learning models. In this talk we will discribe our analysis of the machine learning models, the different language patterns that distinguish individuals having mental illnesses from a control group, and the associated privacy concerns.

We use a dataset[1] of Tweets that are collected from users who reported a diagnosis of a mental illnesses on Twitter. Given the self-reported nature of the dataset, it is possible that some of these individuals are actively talking about their mental illness on social media. We investigated if the machine learning models are detecting the active mentions of the mental illness or if they are detecting more complex language patterns. We then conducted a feature analysis by creating feature vectors using word unigrams, part of speech tags and word clusters [4] and used feature importance measures and statistical methods to identify important features. This analysis serves two purposes: to understand the machine learning model, and to discover language patterns that would help in identifying people with mental illnesses. Finally, we conducted a qualitative analysis of the misclassifications to understand the potential causes for the misclassifications.

2. ANALYSIS AND RESULTS

We will briefly describe our analysis approach and our results in this section and will discuss this in detail during the talk.

2.1 Dataset

The dataset [1] contains tweets from three types of users: users who have self-reported a diagnosis of depression, post-traumatic stress disorder (PTSD) and an age and gender-matched control group. A self-reported diagnosis is a tweet that contains a phrase similar to "I was diagnosed with depression" or "I was diagnosed with PTSD". Such tweets were verified manually to remove jokes, quotes, or any other disingenuous tweets. The dataset contains the most recent 3000 tweets from each user. For our analysis, we were interested in two classification tasks: depression vs. control and PTSD vs. control.

2.2 Direct mentions of mental illnesses in tweets

While reading through a random sample of tweets we realized that some users talk about their condition to raise awareness, to build a support network and to offer help to other users with the same condition. Our initial hypothesis was that simple models such as bag of words are picking up these active mentions of mental illnesses. In our dataset, 24% of users who have depression have mentioned the phrase "diagnosed with depression", and 33% from the PTSD set have the phrase "diagnosed with PTSD/P.T.S.D." or a similar phrase. None in the control group have tweeted a similar phrase.

To measure the effect of such direct mentions of mental health-related issues have towards the prediction accuracy, we trained a classifier to predict if a tweet mentions mental health related issues and removed such tweets from the dataset and measured the overall prediction accuracy. We did not observe a significant drop in performance when the mental health related tweets were removed. The AUC for depression vs. control task dropped by 0.5% and the AUC for PTSD vs. control task dropped by 0.6%.

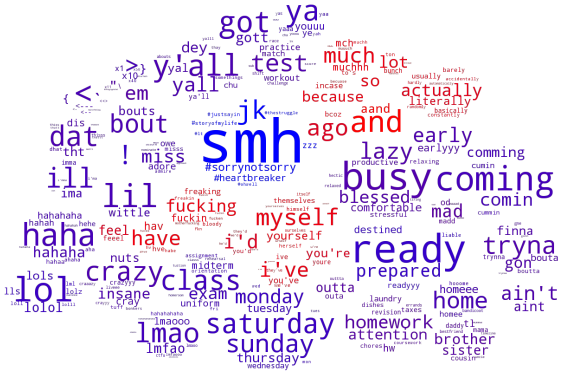
2.3 Feature Analysis

We constructed the following feature sets for our analysis:

- **Bag of Words:** Tf-Idf values of unigrams that are used by more than 1% of the users.
- **POS Tags:** Tf-Idf values of Part of Speech (POS) tag n-grams ($1 \leq n \leq 3$).
- **Word Clusters:** We used a pre-computed set of 1000 clusters[4], where each cluster contains a group of words that are semantically and syntactically related to each other. We computed the Tf-Idf values for each cluster.

We used Information Gain as a mesure of feature importance. To find the features that differentiate the positive and control classes, we extracted features that are statistically significant (Bonferroni-corrected[2] two-tailed p-value of less than 0.05) and has a higher effect size (Cohen's d greater than 0.2).

Figure 1 shows the results from word clusters. Results for the analysis of bag of words features show similar patterns. We discovered several POS Tag n-gram patterns that differentiate individuals with mental illnesses from the control group. Individuals with depression showed higher use of POS tag trigrams with nominal proper noun and possessive verb combination, adjective and adverb tags (Examples: *I'm not, I'm so, I'm pretty sure*), coordinating conjunctions



(a) Depression vs. Control

(b) PTSD vs. Control

Figure 1: Word cloud of significant clusters (Bonferroni corrected $p < 0.05$ and absolute value of Cohen's $d \geq 0.2$). For each cluster the top 10 mostly used tokens are grouped together. The size represents the information gain and the color represents the valency and Cohen's d value - red shades show a higher positive Cohen's d value indicating that the cluster is more frequent in the positive class and a blue shade is associated with the control group.

used together with nouns and pronouns (Examples: *today and, last night and, but I'm not*). Twitter specific discourse markers such as ":" and hashtags were used significantly less by the depressed users. The users with PTSD too showed higher use of conjunctions. Users with PTSD used more n-grams of personal pronouns and past tense and past participle verbs (Examples: *I was told, I was gonna*). Both positive classes showed higher use of personal pronouns, which is a language pattern that was observed in multiple previous studies [3, 5].

2.4 Misclassification Analysis

To understand the machine learning model in detail and to understand the reasons for misclassifications, we manually analyzed a sample of misclassifications. Some of the false positives for the depression class were due to similar language use exhibiting more self focus, anger, and frustration. Some false positives are due to more undesirable reasons such as tweeting about similar topics of interest that were shared mostly by the positive class (such as music bands and artists, and military related tweets). Therefore, we should be mindful about such false positives and understand the limitations when deploying similar machine learning systems in the real world. Interestingly some of the false negatives for depression were from people who have had depression in the past but are now recovered. The analysis of the PTSD vs control classification task seems to show that the classifier associated some of the military related content with PTSD. To validate this hypothesis and to avoid such biases, individuals in the control class need to be matched more closely to those in the positive class.

3. DISCUSSION

Our results show that even after removing direct mentions of mental illnesses, simple machine learning algorithms were able to predict users suffering from a mental illness with a fair degree of accuracy. This means that users who have not revealed their mental health diagnosis or users who have not been diagnosed could be identified by analyzing their social media postings. On one hand, it allows social media platforms or other responsible parties to provide proactive help to users who potentially have mental illnesses. On the other hand, these results raise some privacy concerns. This is especially true given the current revelations that people's psychological profiles were used to target specific advertisements to them during the 2016 US presidential election. Since social media platforms such as Twitter and Facebook allow advertisers to create target audiences by specifying a list of user identifiers[1], it is possible for someone with malicious or unethical intents to create a target audeince who possibly have depression or PTSD.

4. REFERENCES

[1] G. Coppersmith, M. Dredze, C. Harman, K. Hollingshead, and M. Mitchell. CLPsych 2015 Shared Task: Depression and PTSD on Twitter. *the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 31–39, 2015.

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

[3] N. Mor and J. Winquist. Self-Focused Attention and Negative Affect : A Meta-Analysis. 128(4):638–662, 2002.

[4] O. Owoputi, B. O'Connor, C. Dyer, K. Gimpel, and N. Schneider. Part-of-speech tagging for twitter: Word clusters and other advances. 2012.

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

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