Identifying the Mental Health Status of Twitter Users

Kripa Agarwal
Cornell University
2 W Loop Rd, New York, NY 10044
ka467@cornell.edu

Ashley John
Cornell University
2 W Loop Rd, New York, NY 10044
aaj58@cornell.edu

Abstract

Users on social media can reveal meaningful signs of mental health crisis (depression, PTSD, etc.) through the content they share and create. Detecting mental health crisis through analyzing text can have significant consequences for the users and for platform managers. The goal of this paper is to build a novel model to gather, munge, and analyze text from Twitter to understand what types of words and phrases contribute to signals of mental health crisis through natural language processing. The results garnered from this study hope to inform further discussions around mental health, its stigma, and the responsibility of communities and platforms in identifying and helping users in crisis. All code for this project can be found at this Github repository.

1 Introduction

An estimated 300 million people all over the world suffer from depression (dep, 2018). In the United States, more than 22 million people suffer from Post Traumatic Stress Disorder (pts, 2017). At the same time, the presence of social media in teens' and young adults' lives is constant. Users share updates about their lives, retweet and comment on content they find interesting, and interact with other users all over the world. With this rich set of interactions, it is possible that users who are experiencing mental health struggles can express those struggles on social media, which could create the possibility of classifiers that can identify users whose struggle becomes enough to push them into crisis. Though the numbers are staggering and suggest that mental health is truly a pressing international health issue, due to stigmas, mental health often goes undiagnosed. Therefore, if it is possible to use social media data to identify people who may be suffering from mental illness, there is potential to expand mental health treatment and education. Building on the initial baseline model, this new model uses a Recurrent Neural Network for feature engineering to enrich the Multinomial Logistic Regression Classifier of the baseline.

2 Related Work

Coppersmith et al (2014) (Coppersmith et al., 2014) delve into the goal of identifying users in crisis. Their study creates a diagnosed and control group by gathering publically available data from Twitter between 2008 and 2013 through the Twitter API. The data set is referred to as the CLPPsych 2015 data set and is used to build the model used in this paper. Using linguistic tools such as Language Inquiry Word Count, Language Models, and Pattern of Life Analysis, they created a classifier using Linear Logistic Regression.

Kshirsagar et al (2017) (Kshirsagar et al., 2017) look into the identification of crisis with the goal of early intervention. The team uses data from MITs Koko tool which helps platforms manage crisis, abuse, and bullying. To create a classification model, GloVE embeddings and a Recurrent Neural Network are used to achieve 85% accuracy. The difference between this study and the previous Coppersmith work is that the data is coming from a platform that is explicitly used for crisis and harassment work. For this reason, the corpus of words that are used in this text are more complex because they have a higher percentage of crisis users and are not limited by the Twitter word limit.

Socher et al (2013) (Socher et al., 2013) creates sentiment trees which are a corpus with fully la-

beled parse trees that allows for a complete analysis of the compositional effects of sentiment in language. That is, the trees are able to capture how the sentiment of a sentence changes with each word. They also use a Recurrent Neural model for classification into positive/negative sentiment values on a word level. The team also creates n-grams to evaluate sentiment to create a more comprehensive classification model.

Radford et al (2017) (Radford et al., 2017) similarly uses RNNs to evaluate sentiments of movie reviews. The RNN model in this case creates a character level model to learn features about the text and then feeds the features into a Logistic Regression classifier to label each as positive/negative. The model ends up learning a representation of the words through the characters and is able to classify reviews overall base on their sentiment.

3 Dataset Description

The dataset used for this model comes from the Coppersmith et al paper (Coppersmith et al., 2015) which was collected from Johns Hopkins Computational Linguistics and Clinical Psychology Department. The data was collected via the Twitter API between 2008 and 2013 via the following method:

- Diagnosed Group
  - Search Twitter API for the expression "I was diagnosed with X" where X represents a mental illness such as depression or PTSD
    - For each user with a genuine diagnosis, the 3,000 most recent tweets were collected. Remove users with less than 25 tweets and users with less than 75% of tweets in English
      - Tweets were verified for veracity to ensure that they were genuine diagnoses.
- Control Group
  - Use Twitter API to collect most recent 3,000 tweets from a randomly selected list of Twitter users who opted in to a similar Twitter study
  - Remove users with less than 25 tweets and users with less than 75% of tweets in English

The data collection estimated the age and gender of users and created the final data set to control for these factors, which are known to be contributors to mental illness. The training partition consists of 327 depression users, 246 PTSD users for each age and gender-matched control user for a total of 1,146 users. The test data contains 150 depression users, 150 PTSD users for each age and gender-matched control user for a total of 600 users.

The data collection has some imprecisions. First, it relies on users to be honest about their diagnosis and do so publicly. Given the stigma surrounding mental health, it is difficult to gather honest data since many users are hesitant to share their struggles publically. Given this constraint, it is very possible that the diagnosis group represents a group that is more vocal about their mental illness than others who suffer. Additionally, it is possible that those in the control group also suffer from mental illness but do not share their diagnosis publically.

All data was anonymized to respect the users privacy and to uphold the protocol of the the Institutional Review Board (IRB).

In total, 2,292 users were used to train the model for the baseline and approximately, 2 million tweets were used to train the model for the RNN model.

4 Model Description

4.1 Features

The following data processing steps were completed on the text of the tweets:

- Remove URLs, empty rows, special characters, stopwords, numbers, and user names
- Tokenize tweets

For the baseline model, a group of features were added based on additional data collected from the Twitter API: average sentiment of tweets, favorite count, count of the number of first person words, retweet count, count of the number of words identified by Tsguawa et al paper(Tsugawa et al., 2013) as highly correlated with depression, number of uppercase letters, number of exclaimation points, number of question marks, number of ellipses. The additional features were derived for each tweet and then aggregated for each user.

Finally, the dataset used in this analysis has the following labels for the classifier:

- Control: no public announcement of mental illness
- Depression: user had at least one tweet containing ”I was just diagnosed with depression”
- PTSD: user had at least one tweet containing ”I was just diagnosed with PTSD”

The classification labels were encoded as: Control - 0, Depression - 1, PTSD - 2.

4.2 Classification Models

The baseline model sought to classify each user into one of the three classes.

The new model dives deeper to classify each tweet individually. The tweet data contains approximately 2 million training tweets and 1 million testing tweets.

4.2.1 Baseline

As a baseline, a Multinomial Logistic Regression model was used to classify each user into one of the three aforementioned categories. Multinomial Logistic Regression was chosen for this baseline because it assigns a weight to each of the features. For this reason, it provides a useful framework for later exploration to understand which words are most important in classifying a user. Multinomial Logistic Regression also provides a low cost computationally for the baseline analysis given the many features and tweets.

4.2.2 Recurrent Neural Model - Word Level with GloVE Embeddings

Building on the baseline model, an RNN as Language Model was constructed on the word level. For each tweet, Stanfords Twitter GloVe embeddings (Pennington et al., 2014) were used to get word embeddings for each word in the tweet. First, a vocabulary and an embeddings matrix was constructed from the Twitter Glove Embeddings. Next, all the tweets were converted to their word to index format.The tweets were fixed to a maximum length.These tweets were then batched and sent to RNN Model which applied the pre-trained embedding on the words converting them to 25 dimension embedding. The embedded version of the word was sent through the GRU layers of RNN, outputting the hidden state from each word in the tweet. The loss was propagated backward via Cross-Entropy loss between the predicted word of RNN and the actual next word.The final hidden state was recorded to capture a representation of all words in the tweet in the word embeddings dimensions.These final hidden states served as feature matrix which were then passed into the Multinomial Logistic Regression classifier from the baseline model to label each word in the tweet with one of the three categories.

4.2.3 Recurrent Neural Model - Character Level

Next, an RNN was constructed on the character level, For each tweet, the characters were tokenize and encoded into their ASCII integer representations. The encoded version of the tweet was sent into an RNN, outputting the hidden state from each character in the tweet. The final hidden state was recorded to capture a representation of all characters in the tweet. The loss was propagated backward via Cross Entropy loss between the hidden state and the next character in tweet. These hidden states were then passed into a Multinomial Logistic Regression classifier to label each character in the tweet with one of the three categories (Figure 1).

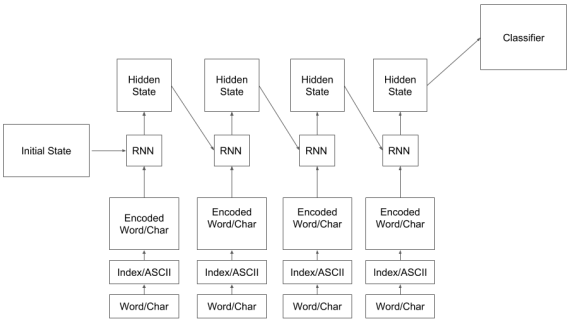

Figure 1: RNN Architecture

5 Experimental Setup

A Gated Recurrent Unit (GRU) is used over a Vanilla RNN since a GRU is able to capture the long term dependencies between words in tweets better than a Vanilla RNN. Vanilla RNNs suffer from the problem of vanishing gradients which prevents the model from learning long-term dependencies.GRU addresses this problem by providing an update gate and reset gate. These two vectors control what information is passed to output and next hidden state.

All work was done in python using packages

such as pytorch, scikitlearn, matplotlib, TextBlob, nltk, scipy, seaborn, and pandas.

To improve the model, the following experiments were conducted:

1. CountVectorizer vs tf-idf: CountVectorizer produces the counts of tokens used in vocabulary as features just like tf-idf. However, tf-idf takes care of the relative importance of the term to adjust for frequently used words.

2. Naive Bayes vs Multinomial Logistic Regression: Naive Bayes gave an accuracy of 48% whereas Multinomial Logistic Regression was 68%. Naive Bayes assumes independence of features and these features are not independent. Multinomial Logistic regression also creates discriminative features whereas Naive Bayes takes a generative approach.

3. RNN on a word level with embeddings vs. word to index: Cross Entropy loss is used with RNN word2index encoding which ensures a higher accuracy as compared to MSE loss in RNN with direct embeddings. Cross Entropy loss is preferred for classification whereas MSE works better for regression. With MSE when comparing the output embedding with the next word embedding, the loss is nearly random since there are very low matches of output word embeddings with the real embeddings.

4. RNN on a character level: for a more granular analysis, an RNN on an character level was created. The character level RNN encodes each character in the tweet with ASCII integer encodings and allows again for the use of Cross Entropy Loss. The character embeddings can learn deeper patterns between words in a tweet which gives a deeper understanding of how words relate to each label.

5. Batching: for the purpose of faster and more accurate training, a batching function was created. The batching function sends in n number of data points and puts them through the RNN model together. Batches of size n=100, n=250, and n=1000 were all tested.

6. Regularization: to combat the potential for overfitting, regularization through additional GRU layers and dropout was used. An extra GRU layer, for a total of two, was added and a dropout layer was induced. Given that there is a lot of noise in the data with each user having a large variety of tweets and language, it is important to limit the noise that the model learns so it does not get overfitted for testing.

7. For each model, 5-fold cross validation was calculated in order to evaluate each model before testing.

6 Results and Analysis

Table 1 shows the accuracies for each of the experiments.

Model	Accuracy
Multinomial Logistic Regression Baseline	0.68
RNN Character Level	0.33
RNN Word Level with GLoVE Embeddings	0.45
RNN Word Level with GLoVE Embeddings and Batch Size n=100	0.48
RNN Word Level with GLoVE Embeddings and Batch Size n=250	0.48
RNN Word Level with GLoVE Embeddings and Batch Size n=1000	0.49
RNN Word Level with GLoVE Embeddings and Batch Size n=100 and 2 GRU Layers + Regularization	0.49

Table 1: Testing Accuracies

The RNN produced much lower accuracies than the original baseline multinomial logistic regression. With the best RNN word level classifier, most tweets were classified as Control or PTSD, but rarely ever Depression. Both the Multinomial Logistic Regression baseline model (Figure 5) and the best RNN model (Figure 6) had difficulty classifying users with mental illness. However, the batching significantly improved the accuracy of the model. With batches, we were able to train for more epochs increasing our accuracy. The batches also allows the loss to converge faster, reducing the training time by approximately 50%.

Although the baseline model solved for user classification and not tweet classifications, the analysis of that model helps understand why the RNN model struggles with the nuances of the text and its labels.

As shown in Figure 2, for control users, the of acronyms like ”smh” or ”lmao” related to happiness are of high importance. Another critical observation is that the probability of all the top 20 words in control set has similar probabilities unlike PTSD and Depression Users, suggesting that users who do not suffer from mental illness have a diverse vocabulary such that no single word is a significant indicator of good mental health.

As shown in Figure 3 , for Depression, the words ”anxiety,” ”f*cking,” ”suicidal,” ”idk,” and ”depression” had a significant impact, but there

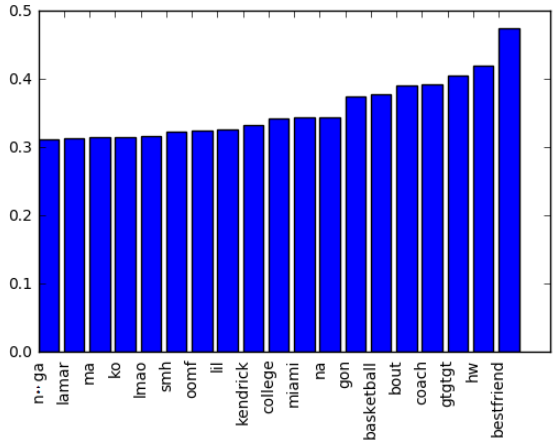

Figure 2: Highest Weighted Words - Control Group

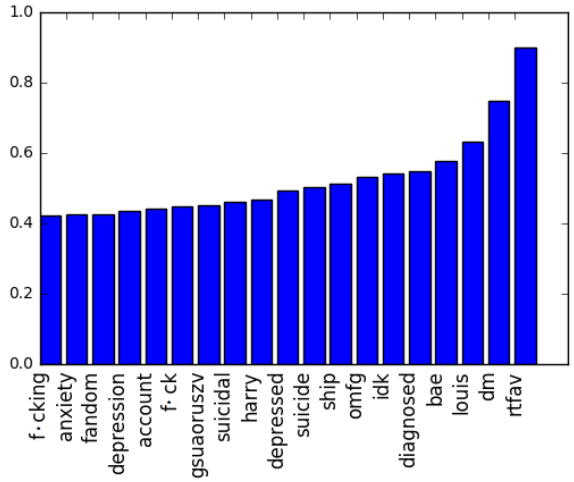

Figure 3: Highest Weighted Words - Depressed Group

were a host of other words totally unrelated to mental illness.

Finally, in Figure 4, words like like "trauma," "rape," "abuse," "military," and "violence" are significant to the label. These set of words aligns clearly with the causes and symptoms of PTSD. It is telling that these words have the highest importance among any of the groups which indicates that they are often used exclusively in talking about PTSD.

The drop in accuracy between the baseline and RNN could be attributed to any combination of the following reasons. First, the initial baseline model using Multinomial Logistic Regression contained additional features on each user on top of a Vectorized and/or TF-IDF representation of each tweet.

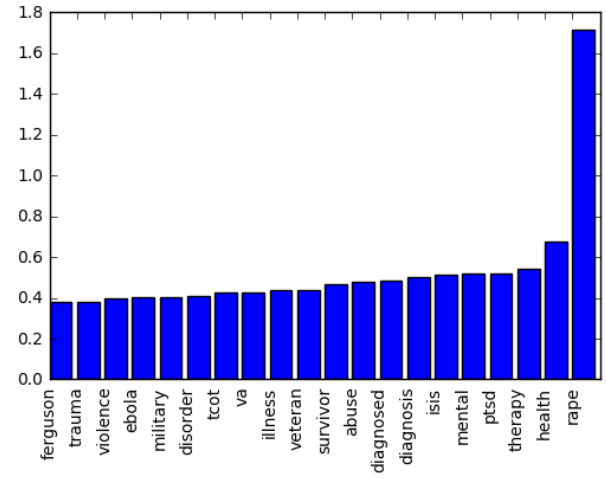

Figure 4: Highest Weighted Words - PTSD Group

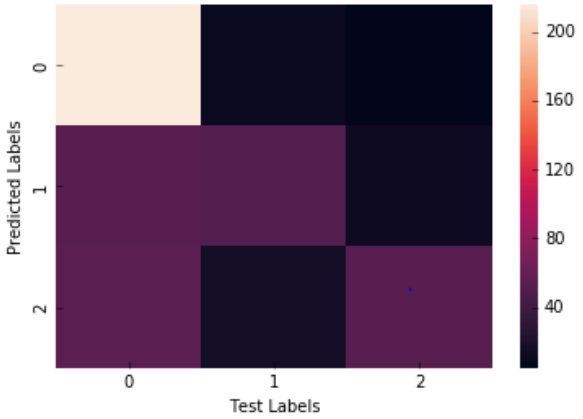

Figure 5: Confusion Matrix - Baseline Model

In the RNN model, these features were not included, thus the model knows nothing about the history of the user. Second, the vocabulary across each of the three classes has a tremendous amount of overlap. The words that people use to describe depression are very similar to the words that people use to describe PTSD. To gain an understanding of how each of the words and characters were being classified, a visualization of each class was color coded such that a tweet is represented in color by each of the three classes. Control was colored green, depression red, and PTSD blue (Figure 7). The content matches the important keywords for each of the classes that was identified in the baseline model.

Finally, the structure of the data set does not lend itself to such granular classification models. The data set contains labels for each of the users,

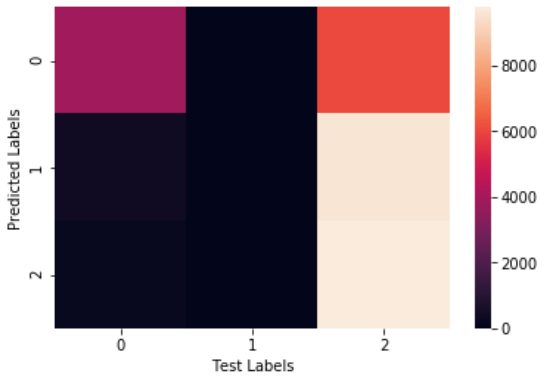

Figure 6: Confusion Matrix - RNN Model

Figure 7: Example Classifications of Selected Tweet shown Control, Depressed, and PTSD respectively

but a user who is depressed does not necessarily always write "Depressed" tweets. Although this data set provides a large sample of users who are depressed or suffering from PTSD, the amount of tweets in the 2 million rows that are explicitly about mental illness is much less. As shown in two example outputs (Figure 8), the content of tweets for each user varies significantly. For example, the first tweet in the visualization of Figure 8 is from a Control group user. However, it is clear from the content and language that they are expressing behavior representing mental illness. Therefore, the RNN model classifies it as PTSD. On the flip side, the second tweet in Figure 8 comes from a PTSD labeled user, but is entirely neutral, which the model again recognizes. Again, the issue here is with the data. The model is able to create a useful language model, but the labeling of the data summarized too much to be useful.

The ideal solution would be to collect data that users report on Twitter as "contemplating self-harm" which would be a clear classification of specific tweets as containing mental health crisis. Unfortunately but understandably, Twitter does not release such information and no current research has been able to acquire it.

Although the model was trained well and recognized labels on a tweet level, the accuracy of the model was inhibited by the data. In the future, a more granular data set could greatly improve the results and hopefully help digital platforms support mentally ill users.

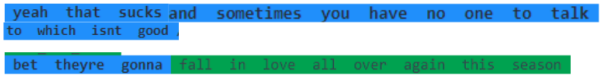

Figure 8: Example Classifications of Selected Tweet shown Control and PTSD respectively

References

2017. Post-traumatic stress disorder (ptsd). https://www.nimh.nih.gov/health/statistics/post-traumatic-stress-disorder-ptsd.shtml.

2018. Depression. http://www.who.int/mediacentre/factsheets/fs369/en/.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in twitter. *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality* https://doi.org/10.3115/v1/w14-3207.

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 2015. Clpsych 2015 shared task: Depression and ptsd on twitter. *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality* https://doi.org/10.3115/v1/w15-1204.

Rohan Kshirsagar, Robert R. Morris, and Sam Bowman. 2017. Detecting and explaining crisis. *CoRR* abs/1705.09585. http://arxiv.org/abs/1705.09585.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*. pages 1532–1543. http://www.aclweb.org/anthology/D14-1162.

Alec Radford, Rafal Jozefowicz, and Ilya Sutskever. 2017. Learning to generate reviews and discovering sentiment .

Richard Socher, Alex Perelygin, Jean Y. Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank.

Sho Tsugawa, Yukiko Mogi, Yusuke Kikuchi, Fumio Kishino, Kazuyuki Fujita, Yuichi Itoh, and Hiroyuki Ohsaki. 2013. On estimating depressive tendencies of twitter users utilizing their tweet data. *2013 IEEE Virtual Reality (VR)* https://doi.org/10.1109/vr.2013.6549431.