Depression Detection Through User Behavior on Twitter

Jamie Yu
jky32

Yunie Mao
ym224

Abstract

Depression, affecting 350 million people worldwide, is a significant public health burden. Despite being a relatively common mental illness, depression is often under-diagnosed, under-reported, and under-treated. One possible way to address this issue is through automatic screening of individuals for depression. This study explored the possibility of automatically detecting individuals with depression through their behavior on Twitter. We selected a group of features most relevant to depression to construct training data for our machine learning model. Then, we explored different machine learning models to test the predictive performance of our model. While much work remains to be done before such a tool could be implemented, our findings demonstrate the feasibility of discriminating depressed from non-depressed users leveraging Twitter data.

1 Introduction

As technology continues to advance, it simultaneously becomes more integrated into our lives and proximal to our behavior. The rapid dissemination of Internet usage via computers and mobile devices provides a large digital footprint that can be mined for various purposes. While many companies in industry use this data on focused ad analytics, an increasing number of researchers are mining the data for potential health care applications. The maturation of Internet technologies and the personal data that it provides has created an unprecedented level of opportunity to develop novel methods to measure, understand, and improve health[2]. In particular, depression - affecting 350 million people worldwide - has emerged as the leading mental health condition of interest amongst computational social scientists as it is a relatively common mental illness and can potentially influence a range of observable behaviors. In addition as the leading cause of disability[14], it is estimated that in the United States, depression costs approximately $50 billion in lost earnings alone in 2010[10].

Despite its ubiquity, under-diagnosis and under-reporting of depression remains a persistent problem: it was found that nearly half(45%) of all major cases of depression in a metropolitan area were undiagnosed[5]. A portion of the considerable public health burden of depression can be lifted by developing technologies to facilitate early screening and diagnosis.

2 Related Work

Depression can potentially influence a range of observable behaviors, which can in turn be measured and utilized to develop screening methods. In 1981, research performed by Bucci and Freedman [3] found that depressed individuals tended to use more first person singular pronouns than those that were not depressed. Since then, many other studies have arisen analyzing the linguistic style of individuals struggling with depression.

With the advent of social media, a number of research studies have utilized digital linguistic data for the purposes of human behavior modeling. Notably, Professor Tanzeem Choudhury of Cornell University has launched several studies exploring the predictive capabilities of machine learning algorithms to detect the onset of mental illnesses. In particular, Choudhury combined both linguistic data from social networks and behavioral data - sleeping cycles, conversation patterns, tone of voice - collected via a user's mobile device to develop the training data for machine learning models[11].

Computational models harnessing data from public forums and textual data from Twitter have all been used to predict the emergence of depression. In a study done by Dodds researching mental illness detection using Twitter data, resulting models both successfully discriminated between depressed and healthy textual content, and compared favorably to general physicians average success rates in diagnosing depression. The model's precision rate had 1 false positive for every 10 depression diagnoses which sharply contrasted with general practitioners from Mitchell et al. diagnosing false positives in more than half of all patients. The results suggest that onset of depression may be detectable from Twitter data several months prior to clinical diagnosis[5].

In a related study also utilizing data from Twitter, researchers from the Georgia Institute of Technology and Zucker Hillside Hospital, Psychiatry Research, aimed to track linguistic markers before and after users self-disclosed mental illnesses on social media[6]. Results showed a large change in affective, cognitive, and linguistic style attributes, and social/personal concerns after disclosing mental illnesses online. In particular, differences in pronoun usage indicated reduced self-attentional focus and lowered social interactivity.

In an exploratory study performed by researchers from various public health institutions in Taipei, Taiwan, researchers developed an Android-based app to capture emotional states and mobile phone usage patterns to predict negative emotions in users. Observed data sources included call logs and history of app usage. Several machine learning algorithms and multiple feature selections were utilized to improve predictability of the models. In total, the average rate of successful detections was 86.17%. Compared with the predictive accuracy of multiple linear regressions (63%) and general guessing (77%), the classifier was substantially more accurate in detecting negative emotions. Thus, the study provided preliminary evidence that it is possible to predict negative emotions via mobile phone usage patterns with substantial accuracy[7].

3 Improved Baseline Model Description

The baseline model is a classification model used to classify the binary categories of depressed or not-depressed amongst users. Thus, we used and tested machine learning methods that perform effectively in classification tasks: Naive Bayes, SVM, kNN, and Random Forest.

The predictive power of our machine learning model comes through assembling a feature matrix that can most effectively differentiate between positive (with depression) users and control (without depression) users. Therefore, our original baseline consisted of a set of linguistic and behavioral features we found statistically relevant in its ability to differentiate between the two groups.

To improve the predictive ability for the improved baseline we explored different feature matrices and ultimately implemented 2 different types:

3.1 Classification Model based on Sentiment Analysis

We implemented a bag of words(BoW), or a sparse vector of the occurrence counts of words, and utilized that as our feature matrix. We were motivated by the fact that a BoW model is a simplistic way to represent the textual data and often used in classification tasks. To modify the BoW so it contained features that would differentiate between the user and control group we applied the following steps to the model:

- **Polarity** Choudhury et al.[12] found that depressed users have a higher negative sentiment in their linguistic tone. Therefore, for each user, we selected tweets with extremely negative(polarity less than 0.7) or positive(polarity greater than 0.7) sentiments to select for the most relevant or charged language.

- **Occurrence** We built a dictionary of the most frequently used negative and positive words (words with occurrences greater than 50) based on all extremely negative and positive user tweets. We filtered by occurrence so we could create a dictionary of the words that would maximize signal and minimize noise.

3.2 Classification Model based on Tweet Metadata

To improve upon our baseline model, we removed features that repeated the same data: we removed 'Average User-tags per Day', 'Average Unique User-tags', and 'Average Unique User-tags per

Day' as we already had the feature 'Average Usertags'. We also added an additional linguistic feature, 'Absolutist Terms', to add to our feature matrix. The following is a final list of behavioral and linguistic features:

3.2.1 Behavioral Features

We observed how users interacted with Twitter and the metadata of the users. De Choudhury et al found that depressed users have lower levels of social interaction[12]. Besides the Insomnia Index, behavioral features measured the level of social interaction with other users on Twitter.

- **Insomnia Index**. De Choudhury et al. found that depressed users tend to post more during the evening than non-depressed users[12]. As a result, they created an "insomnia index" to measure how often a users post at night. In their paper, the "insomnia index" was defined as the normalized difference between the number of posts made during a night window (9PM - 6AM) and a day window (6:01AM-8:59PM) in a users local time [12]. We adopt the same definition for the insomnia index as a feature.

- **Re-tweet Rate**. Rate at which a user shares on their own wall, a post that another user had made. This feature calculates what percent of ones post are re-tweets.

- **Average Usertags**. Users can directly tag another user in their tweets through the symbol @usertag. This feature measures the average user-tags per tweet.

3.2.2 Linguistic Features

- **Average Sentiment**. Choudhury et al.[12] found that depressed users have a higher negative sentiment rating in their linguistic tone. This feature measures the average sentiment score per tweet by utilizing the sentiment.polarity feature in the TextBlob python package.

- **First Person**. Based on research by Bucci[3], it has been statistically proven that depressed individuals use a higher rate of first person pronouns. This feature observes the average number of times first person pronouns were used per tweet.

- **Tsugawa Terms**. In their 2013 paper[13], Tsugawa et al. identified a list of ten phrases whose rate of usage had a higher correlation with depression. The ten phrases are as follows (in order of magnitude of correlation, from greatest to least): even if, low fever, very, workplace, hopeless, disappear, too much, sickness, bad, hospital. This feature measures as the average number of occurrences of these ten terms per tweet.

- **Absolutist Terms**. Al-Mosaiwi et al. found that depressed users tend to use more absolutist terms in their language[1]. Absolutist thinking had a statistically higher correlation with depressed or suicidal individuals. A sampling of the absolutist phrases are as follows: always, definitely, constantly, everyone, everything, etc. This feature measures the average number of occurrences of absolutist terms per tweet.

- **URL**. This feature measure the proportion of tweets than contain a URL

- **Number of Words per Tweet**. This feature measures the average number of words per tweet.

- **Number of Characters per Tweet**. This feature measures the average number of characters per tweet.

The complete set of code implemented in Jupyter Notebook can be found at: `https://github.com/ym224/TwitterDepressionDetection`

4 Dataset Description

In our baseline model we manually collected the tweets of 60 depressed and 60 non-depressed users using the Twitter API. For our improved baseline model, we utilized a clinical dataset that had 561 depressed users and 561 non-depressed users that were age, and gender matched. Researchers gathered representative users for multiple mental illnesses by querying for key phrases such as

> "I was diagnosed with X."

Thus the Twitter database was queried with the phrase

> ”I was diagnosed with depression/major depressive disorder/unipolar disorder today.”

The 561 depressed users were gathered who were verified by a system of human evaluation. Every member of the depressed users was matched with an age, and gender counterpart in the control group. While some members of the control group will suffer from depression, assuming a rate of depression equal to the rate found in the general population, that would mean that the level of contamination makes up a minority of the data. Thus no attempt was made to remove the contaminated data - however, the initial statement of depression (”I was diagnosed with X”) was removed bias from the dataset. For each user, up to their most recent 3000 public tweets were included in the dataset, as that was the limit set by the Twitter API. In total our dataset comprised of 1122 users, and approximately 3.5 million tweets.

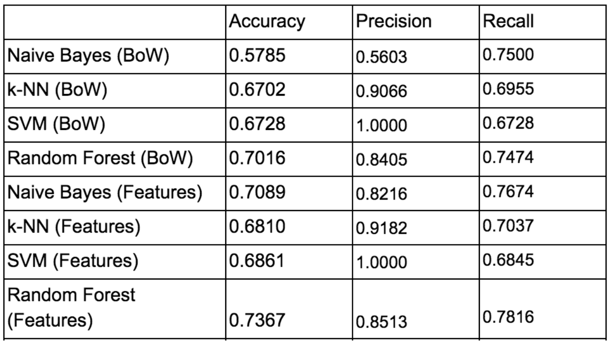
5 Experimental Setup

As discussed in the Related Work section, there has been research to detect depression in Twitter over the last 5 years. This work builds on previous research while exploring different combinations of relevant features on a different dataset. First, a group of depressed and non-depressed Twitter users is identified, and the tweets of the users are collected. Next, a set of features is extracted from these users. We performed preprocessing techniques such as lemmatization and removing of stop-words to standardize the tweets. The data is then split into training(2/3 of total data) and testing data(1/3 of total data). Finally, the machine learning classifier is tested in its ability to identify depressed and non-depressed users (or tweets).

6 Results and Analysis

We fitted our two sets of training data(BoW and feature matrix based on twitter metadata) using the machine learning classifiers - Naive Bayes, Support Vector Machine, K-Nearest Neighbors, and Random Forest. Using the models, we predicted and calculated the mean accuracy on the test data and labels.

Table-1 shows the performance of our models:



	Accuracy	Precision	Recall
Naive Bayes (BoW)	0.5785	0.5603	0.7500
k-NN (BoW)	0.6702	0.9066	0.6955
SVM (BoW)	0.6728	1.0000	0.6728
Random Forest (BoW)	0.7016	0.8405	0.7474
Naive Bayes (Features)	0.7089	0.8216	0.7674
k-NN (Features)	0.6810	0.9182	0.7037
SVM (Features)	0.6861	1.0000	0.6845
Random Forest (Features)	0.7367	0.8513	0.7816

In general, models trained with the feature matrix performed better than models trained on the BoW as evidenced by the higher accuracy rates. For the K-Nearest Neighbors classifier, we experimented with using varying number of neighbors and found the optimal parameter at 9. For the Random Forest classifier, we experimented with using varying number of trees in the forest and arrived at the optimal parameter of 12 trees. Overall, the Random Forest classifier trained on the feature matrix outperformed the other classifiers with an accuracy rate of 73.6%.

In addition, we computed and plotted the confusion matrix for each classifier to find the precision and recall for each model. From the confusion matrices, we can tell that the True Positive rates are the highest for models fitted using Naive Bayes and Random Forest. The True Negative rates are the highest for models fitted using SVM and Random Forest. Amongst the models, Naive Bayes had the highest False Negative rate, and SVM had the highest False Positive rate. Both the K-NN and Random Forest models performed the most consistently when predicting depressed and non-depressed users.

With an accuracy rate greater than the general practitioners from Mitchell et al.[9], who diagnosed false positives in more than half of all patients, our work provides additional evidence that it may indeed be feasible to identify individuals struggling with depression through their behavior on Twitter.

References

1. M. Al-Mosaiwi, T. Johnstone. In an absolute state: elevated use of absolutist words is a marker specific to anxiety, depression and suicidal ideation. *Clinical Psychological Science*. 2018

2. Aung MH, Matthews M, Choudhury T. Sensing behavioral symptoms of mental health and delivering personalized interventions using mobile technologies. *Depression and Anxiety*. 2017; 34:603-609. ¡https://doi.org/10.1002/da.22646¿

3. W. Bucci and N. Freedman, ”The language of depression,” *Bulletin of the Menninger Clinic*, vol.45, no.4, p. 334, 1981.

4. G. Coppersmith, C. Harman, and M. Dredze, ”Measuring post traumatic stress disorder in twitter.” *ICWSM*, 2014.

5. Danforth, C.M., & Reece, A.G. (2017). Instagram photos reveal predictive markers of depression. *EPJ Data Science*, 6, 1-12.

6. Dodds, P.S., Danforth, C.M., Lix, K.L., Langer, E.J., Reece, A.G., & Reagan, A.J. (2017). Forecasting the onset and course of mental illness with Twitter data. *Scientific Reports*.

7. F Rizvi, Asra & L Birnbaum, Michael & M Kane, John & De Choudhury, Munmun & Ernala, Sindhu Kiranmai. (2017). Linguistic Markers Indicating Therapeutic Outcomes of Social Media Disclosures of Schizophrenia. Proceedings of the ACM on Human-Computer Interaction. 1. . 10.1145/3134678.

8. Hung, G. C.-L., Yang, P.-C., Chang, C.-C., Chiang, J.-H., & Chen, Y.-Y. (2016). Predicting Negative Emotions Based on Mobile Phone Usage Patterns: An Exploratory Study. *JMIR Research Protocols*, 5(3), e160. ¡http://doi.org/10.2196/resprot.5551¿

9. Lin, L. Yi, Sidani, J. E., Shensa, A., Radovic, A., Miller, E., Colditz, J. B., ... Primack, B. A. (2016). Association between Social Media Use and Depression among U.S. Young Adults. *Depression and Anxiety*, 33(4), 323-331. ¡http://doi.org/10.1002/da.22466¿

10. Mitchell, A. J., Vaze, A. & Rao, S. Clinical diagnosis of depression in primary care: a meta-analysis. *The Lancet* 374(9690), 609619 (2009).

11. P.E. Greenberg, A.-A. Fournier, T. Sisitsky, C.T. Pike, and R.C. Kessler. (2005 and 2010). ”The economic burden of adults with major depressive disorder in the united states,” *The Journal of clinical psychiatry*, vol. 76, no. 2, pp. 1-478, 2015.

12. T. Choudhury, A. T. Campbell, N. D. Lane, H. Lu, Y. Xu and S. B. Eisenman, ”Exploiting Social Networks for Large-Scale Human Behavior Modeling,” *IEEE Pervasive Computing*, vol. 10, no. , pp. 45-53, 2011. doi:10.1109/MPRV.2011.70

13. S. Tsugawa, Y. Kikuchi, F. Kishino, K. Nakajima, Y. Itoh, and H. Ohsaki, Recognizing depression from twitter activity, in Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. ACM, 2015, pp. 31873196.

14. ”Depression [fact sheet no.369],” 2015. *World Health Organization.*