Signals of Depression in Social Media: A Look into the Feasibility of Automated Depression Detection

Rob Maron, Matthew Peeks, Dade Wood

December 7, 2020

Abstract

Millions of people struggle with depression worldwide. While there are many resources for these people to get help, a large percentage of people who suffer from this illness are never treated because they do not recognize the symptoms in themselves. To assist these people in finding the help they need, many researchers have developed language models that can detect depression in social media posts. If implemented into a social media network, the models could act as an early detector of symptoms and get those who need help to seek it sooner. However, to make this ideal a reality, the predictions need to have a higher level of accuracy than they currently do. The importance of detecting depression early and the prevalence of self-diagnosing means that false negatives and false positives can adversely affect a person's life. As a result, there is very little acceptable margin for error for any application of this system. This project will focus on improving existing social media depression detection models and analyzing the feasibility of such models in order to be one step closer to helping people who suffer from depression.

Throughout this project, we will seek to answer: What do social media posts that indicate signs of depression look like and are there common features amongst the posts can be used to detect signs of depression? Is it possible to reliably and accurately detect signs of depression from social media? Any models created will be measured by the use of standard metrics such as accuracy, precision, and recall.

1 Introduction

Depression is a serious mental illness, causing symptoms ranging from a constant feeling of sadness, trouble sleeping and eating, lack of energy, feelings of worthlessness, and, in extreme cases, suicidal thoughts. Over 264 million people suffer from this disease and, despite the incredibly high statistic, between 76% and 85% of people in low- and middle-income countries never receive treatment for their disorder [1]. One of the major reasons for this troubling percentage is that the diagnosing of depression is an incredibly difficult and labor-intensive task [3]. Automating this task would assist individuals to receive the treatment they need.

One way to automate this task would be to implement a model that monitored everyone's social media posts and detected depressed language in their posts. If the model recognized that a user's post history is showing signs of depression, it could notify the user that they may benefit from seeing a professional about a diagnosis. Models that attempt to achieve this goal on various social media platforms have been developed, but none have been accurate enough to be implemented to a larger scale. This project offers an attempt at making models to detect depression in people based on their social media posts and analyzes how accurate they are.

2 Data Collection

This problem can be simplified to a classification task on categorical data. The two categories are tweets from individuals who suffer from depression and tweets from individuals who do not suffer from depression. To collect this categorical data, tweets were scraped from Twitter using the python package Twint. The tweets from people with depression were found by searching for users who had tweeted that they had been diagnosed with depression. Tweets from people who do not suffer from depression were originally tried to be found by scraping tweets with no restriction, but Twint defaults to popular tweets when not given instruction which would make the data biased on popularity. Instead of this, famous people's and company's twitter accounts had small subsets of their followers' tweets scraped. This allows for a larger subset of Twitter's user base to be represented, as a substantial percentage of Twitter users follow these kinds of accounts, but still restricts the non-depressed data from coming from all of Twitter.

In addition, there is no method to verify the users gathered in this way are not depressed, so the assumption is made that a large majority of tweets scraped would be from non-depressed individuals and that the data incorrectly being classified as non-depressed would not be substantial enough to affect the models' performances. There is a similar issue with the subset of users labelled depressed. Without having users in the data set show proof from a doctor that they have been diagnosed with depression, it cannot be verified that the users truly are depressed so they are only la-

belled as so based on the assumption that they were correctly diagnosed by a professional.

3 Feature Extraction

In order to classify a user as depressed or non-depressed from their tweet history, we must first find attributes of the tweets that characterize differences between each class. The features we chose to try that have been determined to be significant per user in previous research are the number of positive/negative sentiment tweets, word embeddings [7], average time gap between consecutive tweets, average word count in tweets [1], most active time period of all tweets made, usage of 3rd person vs 1st person pronouns, average swear word usage [2], sentiment of emojis used, and count of depression symptom terms used among all tweets [8]. After determining how well they are able to separate the two classes, these features are then used in combination in order to create the machine learning models.

In the final models, we remove the most active time period and the average word count per tweet because they were not significant enough to distinguish the classes and did not affect model accuracy in a positive way. We also remove word embeddings from features as it does not fit well into a matrix with the other features because we are classifying each user instead of each tweet as being depressed or not depressed. The word embeddings would be useful as a separate model on its own but since this has already been done [7] we decided against using it.

Feature	Depressed	Not Depressed
Positive Tweet Proportion	0.4064	0.4578
Negative Tweet Proportion	0.5936	0.5422
Time Gap Between Tweets (Days)	3.7736	17.4546
Word Count Per Tweet	15.2755	11.4952
Swear Words Per Tweet	0.0752	0.0606
Symptom Mentions Per Tweet	0.1216	0.0654
First Person Pronouns Per Tweet	1.012	0.7498
Third Person Pronouns Per Tweet	0.4688	0.345
Most Active Time Period	4-8pm	4-8pm

Figure 1: Feature Table
This table displays the average of the feature stated for each class except in the most active time period where it displays the time period itself instead.

Most of the values in Figure 1 are fairly close for each class but were still determined to be significant by an increase in accuracy of the models. What should be noted, however is that many of these values are different from what we expected from previous works. For example, the time gap between tweets[1] and 3rd person pronoun usage [2] were expected to be larger for depressed individuals but our data showed the opposite. Likely, this difference is caused by utilizing different datasets and the possible biases in our dataset already discussed.

4 Evaluation of Models

We chose two different model types often used for this problem using the features described: random forest [1] and a neural network [7].

	NN Model		RF Model	
Stats	Not Depressed	Depressed	Not Depressed	Depressed
Precision	0.81	0.76	0.85	0.77
Recall	0.75	0.82	0.76	0.85
F1-Score	0.78	0.79	0.8	0.81
Support	309	291	309	291
Accuracy	0.78		0.81	

Figure 2: Evaluation Metrics

We use standard classification evaluation methods such as accuracy, precision, and recall to determine how well each model does on a subset of testing data set aside. These values are shown in Figure 2 where we can see that the random forest model has a slightly better performance than the neural network. This could be attributed to the fact that many of the features implemented are more behavioral features of the user in how they write their tweets which work well in a decision tree where the features can be split into value ranges that determine the class.

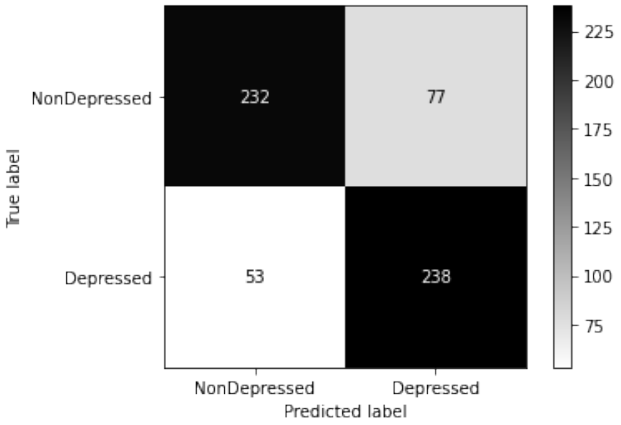

Figure 3: NN Confusion Matrix

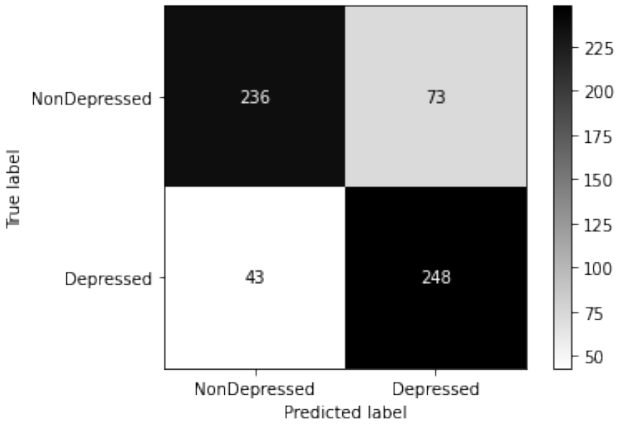

Figure 4: RF Confusion Matrix

For a final part of evaluation, we examine the confusion matrices of the models to determine what each model tends to classify the users. As shown in Figure 2 and Figure 3, both models have a much higher count

of false positives than false negatives. This could be attributed to a number of things but is likely a result of the number of tweets per user, and thus the amount of data, being higher on average for depressed users than for non-depressed users which is a result of our data collection methods.

5 Discussion of Ethics

Attempting to detect depression raises many ethical concerns about privacy, impact on people's lives, and the dilemma of having machines involved in diagnosing a mental illness.

There are two important parts to the issue of privacy. First, because this is a medical problem, any detection of a person showing signs of depression needs to be kept completely private and secure. This means that when the system detects a user as showing symptoms, only the user should be notified and in way that keeps this information private. And second, the nature of the classification problem leads itself to needing lots of data in order to properly classify one way or the other. This requires a deep analysis of a large amount of user data. However, people may not want their data being used in this way, so in order for the model to be run on the platform, each user must acknowledge and agree to their data being gathered and analyzed.

While this technology is meant to help people, it could very easily do the opposite. For example, if someone gets detected as a false positive, they could start to believe that they have depression even before they confirm it and get diagnosed by a doctor, thus leading them to make different choices than they may usually make as a form of confirmation bias. This means that, for a model to be viable in practical use, it needs to limit the amount of false positives it has as much as possible and focus more on the elimination of false positives than of false negatives.

Finally, there is the issue of having a computer involved in a process of aiding depression diagnoses. Mental illness is expressed differently for every individual, making it so hard to diagnose that many professionals struggle with the task [7]. With there being so much mystery still left in how the human brain functions and how mental illnesses come about, people may question if it is even right to hand over tasks involving complex human emotion and behavior to a machine. The only way to address this concern is for psychologists to develop a better sense of what depression is and what commonalities it has across all people so that models such as these are able to become accurate enough that people are comfortable utilizing them.

6 Conclusion

From the results, it is surprising to see how even simple features put into the right model can cause a relatively high accuracy. While the highest accuracy of about 80% may not seem significant, especially when considering how an implementation of this system would certainly need a much higher accuracy, it is important to note how simple the systems used in this project were. This 80% can be seen almost as a baseline, and future works can begin to achieve that final 20%. To push the accuracy up to a level where the model could be applied, the first step will be using more features of the tweets themselves. Because of the limited amount of time that could be dedicated to this project, only more basic features could be utilized. For example, emoji sentiment was desired to be a feature considered by our models, but that would require a dataset of all emojis and their corresponding sentiment, which would be very time consuming to produce. Once more features were considered by a model, the choice of model will likely make more of a difference.

This paper only considers two models, the random forest and the multi-layer perceptron neural network, but other models such as a linear support vector model, logistic regression, naïve bayes, etc. could potentially yield better results. In addition, while scikit-learn's multi-layer perceptron was utilized in this paper, a more fine-tuned neural net could potentially handle this problem better than the simpler models used.

This paper aimed to analyze the viability of automated depression detection in tweets. The fact that these results were achieved with such simple systems shows that current natural language processing techniques are potentially enough to create a model accurate enough for actual application, provided that enough time was spent crafting the features extracted and the model as well as gathering a sufficient amount of less biased data.

7 Acknowledgements

Rob Maron for data collection from Twitter and word embedding analysis. Matthew Peeks for report writing/reviewing and presentation proofreading. Dade Wood for feature extraction and model creation, report writing/reviewing, and presentation writing.

References

[1] Fidel Cacheda et al. "Early Detection of Depression: Social Network Analysis and Random Forest Techniques". In: *J Med Internet Res* 21.6 (June 2019), e12554. ISSN: 1438-8871. DOI: 10.2196/12554. URL: http://www.ncbi.nlm.nih.gov/pubmed/31199323.

[2] Munmun De Choudhury et al. "Predicting Depression via Social Media". In: AAAI, July 2013. URL: https://www.microsoft.com/en-us/research/publication/predicting-depression-via-social-media/.

[3] Sharath Chandra Guntuku et al. "Detecting depression and mental illness on social media: an integrative review". In: *Current Opinion in Behavioral Sciences* 18 (2017). Big data in the behavioural sciences, pp. 43–49. ISSN: 2352-1546. DOI: https://doi.org/10.1016/j.cobeha.2017.07.005. URL: http://www.sciencedirect.com/science/article/pii/S2352154617300384.

[4] Jana M. Havigerová et al. "Text-Based Detection of the Risk of Depression". In: *Frontiers in Psychology* 10 (2019), p. 513. ISSN: 1664-1078. DOI: 10.3389/fpsyg.2019.00513. URL: https://www.frontiersin.org/article/10.3389/fpsyg.2019.00513.

[5] Huijie Lin et al. "Detecting Stress Based on Social Interactions in Social Networks". In: *IEEE Transactions on Knowledge and Data Engineering* PP (Mar. 2017), pp. 1–1. DOI: 10.1109/TKDE.2017.2686382.

[6] Adil Rajput and Samara Ahmed. *Making a Case for Social Media Corpus for Detecting Depression*. 2019. arXiv: 1902.00702 [cs.CL].

[7] Marcel Trotzek, Sven Koitka, and Christoph M. Friedrich. "Utilizing Neural Networks and Linguistic Metadata for Early Detection of Depression Indications in Text Sequences". In: *IEEE Transactions on Knowledge and Data Engineering* 32.3 (Mar. 2020), pp. 588–601. ISSN: 2326-3865. DOI: 10.1109/tkde.2018.2885515. URL: http://dx.doi.org/10.1109/TKDE.2018.2885515.

[8] Hamad Zogan et al. *Depression Detection with Multi-Modalities Using a Hybrid Deep Learning Model on Social Media*. 2020. arXiv: 2007.02847 [cs.IR].