Depression Detection: The Impact a User's Mental Health Can Have on Their Online Behavior on Reddit

Huy Tran
New York University
htn279@nyu.edu

Abstract

The percentage of people who experience depressive symptoms has increased with the start of the pandemic, and we know that, if left unchecked, depression can have extremely severe consequences on the lives of individuals all across the world. As such, we decided we wanted to find a way to catch out people who were displaying suicidal tendencies on a large social media platform that's known for creating hive-mind mentalities, which can be incredibly dangerous in situations like this—Reddit. So we found a method to detect suicidal users on Reddit. Here's how we did it. First we identified posts outside of r/depression that were likely written by people who were depressed, and then we extracted linguistic patterns to distinguish between the threads written by people who had had depression and people who didn't so that we could shed some more light on this field of research as a whole. We found that a Gradient Boosted Tree at 1000 estimators, a max depth of 5 and a learning rate of 0.2 gives us the most accurate results–furthermore, we will also focus on our F1 score to balance between precision and recall. We aim to share this information with organizations like the National Institute of Mental Health or the Substance Abuse and Mental Health Services Administration in the hopes that it can be used to do further research into the negative impacts that online communities can have on mental health. All in all, we aim for our results to be used as preventative data points that can target specific users in advance, and hopefully work towards decreasing the amount of negative content they see. This is with the intention of decreasing suicide rates seen in Reddit users and improving their overall quality of life in the small ways we can.

1 Introduction

Depression is a kind of mental health disorder characterized by persistently low mood and loss of interest in activities–severe depression will cause significant impairment of daily life. At least 2 to 6 percent of people in the world experience depression, and in the US, 71.0% of those people received help, according to data posted by the National Institute of Mental Health. During the pandemic, the rate of anxiety and depression globally increased by 25%. However, not all people seek professional help as soon as they discover symptoms related to depression. 90% of teens and young adults with symptoms of depression said they had searched online for information about mental health issues, according to a survey created by Hopelab, and one of these forums for searching just so happened to be Reddit. To really make a difference it's key to find these people in the early stages of depression and provide them relevant treatment as soon as possible. Our project will focus on finding people with the possibility of depression on Reddit timely before their symptoms deteriorate and lead to more extreme behaviors.

Reddit is a social forum that contains millions of communities of people who share common interests. One may find discussion such as breaking news, TV programs, pets, and makeup techniques on this forum. There are subreddits that allow users to focus on a specific topic in posting content that is voted up or down by relevance and user preference. People may also comment further after other people's posts and create a thread after it, and all their posts are anonymous.

There's a specific community on Reddit called r/depression that provides support for people experiencing depressive symptoms. In r/depression, moderators only allow a post to be submitted if the poster is seeking help with their depression, which provides a unique chance for us to have a cohort of people who are **likely** to be depressed (these people will be referred to as D+ users) without having to label the data ourselves. The reason why we use "likely" is because the posts only have to be examined by the subreddit moderators instead of formal diagnosis, which provides an opportunity for there to be people who are self-diagnosed or aren't telling the truth about their situations.

In our project, we will mainly focus on other communities–instead of r/depression–and try to identify which posts are written by users who are likely depressed. Through our analysis, we produced several models, and identified topics and patterns that will hopefully be of help in identifying the undiagnosed.

2 Dataset

For our dataset, we created a crawler to collect posts from different subreddits based on our own pre-defined rules.

We also ended up using Pushshift to collect our data for this project. We searched through r/depression from 2019 to 2020 to form our D+ cohort, resulting in roughly 50,000 users. We chose this duration because, first of all, this time is the duration of covid virus, which we choose to examine for our project, as stated before. Also, after a preview of the data, we found out that the data in 2021 was two times larger than the 2020 data, which was significantly easier to analyze.

To categorize and analyze the data, we collected each of these users' first month of posting and made these our study months, as it's difficult for us to collect the data across more than one month due to the nature of the data we collected. We then extracted each D+ users' associated month to collect any text-based posts outside r/depression.

To collect our D- users, we took the most popular subreddit any D+ users had posted in and found a user who had posted in that same month and who was not in the D+ cohort. Once we created a list of users, we went through all of their associated months to collect all the text-only posts they submitted. Our data, not including the titles or the comments in the posts, consisted of text only.

After collecting data, we performed some cleaning to exclude outliers such as posts with an inappropriate length, duplicate posts, and posts from other depression related sub (r/SuicideWatch for example). Our D+ cohort was a positive class and D- cohort was a negative class.

Our final corpus contained 102,523 posts, with 49% positive and 51% negative class split, with 15712 D+ users to 14352 D- users.

The reason why we collect our own datasets is that it allows us to construct a cohort that would best help us to identify trends specific to online discussion forums and under how users with depression might seek help on social media. However, our definition of D+ is only defined by having posted in r/depression, not in a literal sense. Also, due to the nature of Pushshift and our limited access to computing resources, our data collection is not perfect and can still be improved in future work.

3 Data Preprocessing

For our data preprocessing, we first prepared the data text by standardizing it to reduce the noise in it, so that we could focus on its content. We then performed standard data cleaning, such as lowercasing the text, expanding its contracts, removing punctuations in the data, and dealing with digits and accents. We removed reddit specific content like “/u”, “/r” and dropped URLs, and brought the number of tokens in our data from 10683291 to 1000281 and from 484243 types to 131082 types. Finally, we split our data into train validation and test sets.

Our next steps included using Grensim’s Phaser [R. Rehurek and P. Sojka. Gensim–python framework for vector space modeling . NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic, 3(2), 2011.] to create n-grams based on our data. We expect these n-grams to occur in at least 0.05% of the train dataset. We finally made 8138 n-grams, with max n = 4.

To conclude, we applied stemming to our data to remove those suffixes of words and lemmatization to convert words into a more standardized format, and finally obtained 4820138 tokens and 118235 types.

4 Exploratory Data Analysis

Below we will present our exploratory data analysis that we performed on our collected dataset. This will include both the raw text and the preprocessed text. It should be noted that some of the posts might contain informal language, which would lead to some of the tokens being skipped–.

	D- Minimum	D- Median	D- Mean	D- Maximum
Raw Text T	13	102	158.2	2432
Raw Text T	21	78	87.93	492
Raw Text S	1	8	8.232	128
Preprocesse	1	52	66.32	50
Preprocesse	1	45	54.12	421

Figure 1: Compilation of D- data

	D+ Minimum	D+ Median	D+ Mean	D+ Maximum
Raw Text T	17	115	159.2	2821
Raw Text T	2	78	88.13	530
Raw Text S	1	7	7.812	472
Preprocesse	1	52	67.68	578
Preprocesse	1	43	53.72	353

Figure 2: Compilation of D+ Data

4.1 Corpus Statistics

As we can see in our table above, even though we’ve defined our crawler to exclusively keep posts with less than 450 tokens, there are still several comments with extremely high token counts. This is most likely due to the nature of token-handling mechanics in Python and the EDA library in R. That being said, we found that these counting issues proved not to create any kind of significant problem after the data was properly preprocessed.

By preprocessing the data, we actually managed to reduce the mean number of tokens by more than 50%, which is relevant, because it was extremely beneficial for the modeling techniques that we ended up using later. Of course, we also had to remove many of the extreme cases to avoid skewing, and when we look;’ in to the distribution, we can see that they share a similar structure without any distinguish characteristics.

4.2 Word Cloud

After preparing and preprocessing the data, we used word cloud feature from LIWC to see if we could find any sort of clear distinction between D+ and D- users. Below, we’ve included the depression cloud that we got.

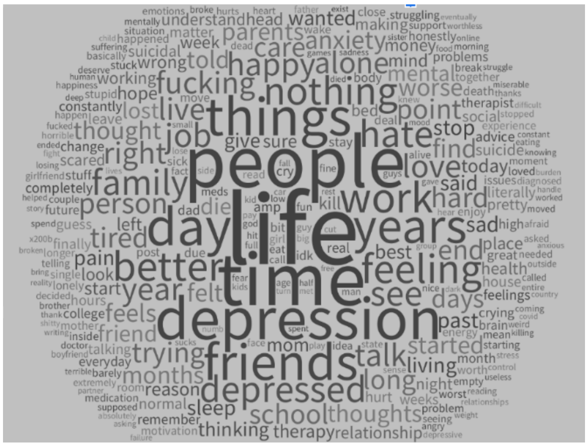

Figure 3: Word cloud with discovered terms

Once we looked at this, we saw that that using text analysis techniques to catch early signs of depression could have quite promising results that could be used an further elaborated upon by other researchers.

5 Analysis

5.1 Word Counting Approaches

First, an introduction to word counting. Word counting is calculating the word frequency of tokenized text, which essentially means finding the exact number of times that a specific token occurs in that text.

So, after being inspired by some prior research in our field, we actually attempted to perform some word counting ourselves. We utilized the Linguistic Inquiry and Word Count's (LIWC's) dictionaries to determine which terms we were going to search for, and ended up identifying similar trends that were observed by Ramirez-Esparza et al. using our technique.

In the cited paper, the author clearly states that D+ users preferred singular pronouns over their plural equivalents. Additionally, they also seem to prefer words with negative connotations more than D- users. From there, we counted the terms from each post that fit into each category, figured out the proportion compared to the length of the posts, and summarized it all in the table below.

Cohort	First-person Singular	First-person Plural	Postive Terms	Negative Terms
D-	5.23% (3.72%)	4.72% (1.12%)	3.21% (2.49%)	3.02% (2.72%)
D+	6.72% (3.82%)	3.80% (0.99%)	3.21% (2.49%)	3.57% (2.95%)
Difference	-1.49% (-0.10%)	0.92% (0.13%)	0% (0%)	-0.55% (-0.23%)

Figure 4: Summarized data from the clouds we created through analysis

5.2 Bag-of-Words

Now, a quick introduction to bag-of-words. Bag-of-words is used to extract features from text so that the information can be properly used when modeling.

Similarly, after standardizing our text, we needed to represent it in a way that could be used for modeling. In a scenario like this, the most common approach is to simply convert each post into a frequency of each term that's present, and after doing that, we ended up with a vector that was the size of the final vocab, and was filled in by the counts of terms that were found in its index, along with zeroes for all of the missing words. A variant of this the is TF-IDF approach, which weights terms by the number of times it appears across a corpus, but we found that our approach was more to-the-point for our purposes.

5.3 Modeling

Finally, we trained a supervised model to predict which posts were written by users who demonstrated depressive tendencies. We attempted several models of classification in our research, and will go over, what we chose, as well as our results, below.

5.3.1 Naïve Bayes

Naive Bayes is essentially a model that's used in classifying–it takes the Bayes theorem, and the unique features are assumed to be independent. What we did was utilize Sci-kit Learn's implementation of Naive Bayes with a BoW vector of each doc. We then learned the estimated probabilities, and classified the out-of-sample posts.

5.3.2 Logistic Regression

TF-IDF is a way to check how frequent a word is in file. This basically interprets the importance of each word to the file that it's in. We ended up utilizing Scikit Learn's implementation of TF -IDF since it comes with a built-in normalizer, and this helped us to come up with commonalities

5.3.3 Decision Tree

Finally, the decision tree, which uses input and output to train models and is used for classification and forecasting. But one thing about a decision tree is that it can be prone to overfitting. To troubleshoot this, we utilized enhanced variants of random forest [L. Breiman. Random forests. Machine learning, 45(1):5–32, 2001] and Gradient Boosted Tree [J. H. Friedman. Greedy function approximation: a gradient boosting machine. Annals of statistics, pages 1189–1232, 2001.] [intro to random forest][intro to gradient boosted tree]. Generally, these are robust to noise and offer competitive results. Although these trees are powerful machine learning techniques in terms of performance, they only give us the "importance" of features. Extra steps are required to understand the effect of said features on actually predicting depression, which we leave for future work.

6 Results

Model	Vector Representation	Accuracy	Precision	Recall	F1
Gradient Bo	BoW + N-grams, Counts	0.68	0.54	0.73	0.61
Random For	BoW + N-grams, Counts	0.64	0.56	0.79	0.66
Logistic Reg	BoW + N-grams, TF-IDF	0.62	0.61	0.54	0.57
Bernoulli Ba	BoW + N-grams, Binary	0.61	0.58	0.6	0.6

Figure 5: Results from modeling

We've reported our results in the table above, and as seen here, despite our efforts, our results are quite lackluster. We continually see accuracies that hits 0.68 at best, and barely make it past 0.60 at worst. We clearly get our best result from the Gradient Boosted tree model at 1000 estimators, a max depth of 5 and learning rate of 0.2.

While accuracy is a fair metric to use, we also want to focus on our F1 score to balance between precision and recall. If we consider this, for us, the highest F1 score is achieved with Random Forest at 500 estimators and a max-depth of 3.

Furthermore, it should be noted that these results don't tell the whole story–we were also able to extract trends within our text that could help distinguish the two cohorts.

7 Model Insights on Text

Due to their nature, we were able to extract some insights from our Naive Bayes and logistic regression-based modeling.

From our Naive Bayes model, we saw that the phrases that are associated with D+ users tend to have content that identifies themselves as being depressed like "my depression" or otherwise identifies themselves with other mental health issues, using terminology including "cutting", "anxiety" and "burned out." One additional finding is that there are a significant amount of terms related to deadlines–and interestingly enough, poetry–that also seem to be correlated with people who are demonstrating aspects of depression as processed by us. On the contrary, terms that related to sports reduced the chance of a post being written by a D+ users.

From our log regression model, terms that increased the likelihood of being written by D+ users included "college," "job," "work," and overall profanity.

8 Related Work

8.1 The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods

This article[12] discussed the linkage between real-world behaviors and actions, and daily word use. It also went over Linguistic Inquiry and Word Count (LIWC), and how it could be used to find meaning in experimental settings, which included emotionality, thinking styles, and individual differences. This overlapped with our project idea–particularly the analysis of these unique settings–and so we used this is inspiration, and used LIWC to process our data through the guides of our three linguistic cues.

8.2 Emoji semantics/pragmatics: investigating commitment and lying

This paper[9] discussed the concept of deceptive emoji use, and how certain emojis could be used in greatly varying ways, with greatly varying intentions. While our data and results weren't actually all that similar, we knew that there would be crossovers in the process, such as the methods that this study uses to determine which emojis were used with the intent to deceive, which is reminiscent of how we decided to determine which key phrases are genuine and which ones are not.

8.3 Deep Learning for Depression Detection of Twitter Users

This paper[10] was about efficient deep neural network architectures that have priorly been used to process language and applying them to unstructured text data extracted from Twitter. This data was then analyzed to attempt to identify mental illnesses through social media platforms. One of our primary struggles did end up being the processing datasets, so we were able to take inspiration from some of the methods outlined here. Of course, while the general task was similar as well, this ended up relating back to our project in terms of strategy identification.

8.4 Expert, Crowdsourced, and Machine Assessment of Suicide Risk via Online Postings

Finally, this paper[11] discussed the relationship between the risk of suicide and online postings on Reddit, and developed a new approach for determining the classifications of "at risk." As such, the method for data collection–referenced by Coppersmith et al. (2014)–ended be similar to our approach. It consisted seeking of statement variations that related back to the original purpose, and manually figuring out which phrases are genuine and which ones are not. In a similar way, our system identified these statement variations, and used them to calculate the depression values. We also used the results as a sort of general baseline comparison for ours.

9 Conclusion

Our work has revealed some present results that begin to signal at success, and creates connections between used phrases and depression, which is what we were aiming for. As of right now, our models tend to predict with 60-70% accuracy, which is far lower than previous works, but we accept that trying to predict depression outside of depression-specific space is a harder challenge. Lastly, we also revealed that people who use Reddit as a space to discuss their hobbies tend to be non-depressive users, which is almost the opposites of users who use Reddit to seek advice on personal relationship and mental health issues, who tend to be marked as depressed more often than not.

To conclude, we once again reiterate the importance of a motivation like ours. Depression comes hand-in-hand with a painfully high mortality rate, so we'd like to do everything we can to make life a little easier for the people who experience it. Hopefully, our approach, although not as fleshed out as previously referenced papers, can be used as a basic starting point for future work, where perhaps a user's entire post history can be analyzed instead of just a single post, as we do here. This could lead to results where we'd catch out earlier stages of depression, before it evolved into something worse.

10 Future Work

Although we collected our data carefully and tried to eliminate all of the disturbing outliers, we acknowledge that there are still imperfections in our work. One example is that we collect data from the year 2021, as the data from that year is 2 times larger than 2020 data, so we cannot make sure that our dataset includes all of the cases of depression throughout the Covid period. Also, since we collected data from each users' postings from the first month, we lose track of their situations in the following months, in which some of those users may have depression symptoms while they do not reveal any of them during the first month. Third, our data is analyzed based on text, while there are many depressed people who express their emotions through other ways, such as emoji and meme pictures. Due to the type of our data, we are not able to detect those users' depression symptoms by now. Finally, as we mentioned above, the priorly referenced trees are extremely powerful machine learning techniques as far as performance goes, but they don't really help us in terms of really understanding and analyzing the effects of these features in actually predicting depression. Instead, it only tells us the so called "importance" of particular features. This is something that requires extra steps are require, and as such, is something that we will leave this to future work.

11 References

Rideout, Victoria, and Susannah Fox. "Digital Health Practices, Social Media Use, and Mental Well-Being Among Teens and Young Adults in the U.S." Hopelab, 2018, https://hopelab.org/reports/pdf/a-national-survey-by-hopelab-and-well-being-trust-2018.pdf.

"Major Depression." National Institute of Mental Health, U.S. Department of Health and Human Services, 2022, https://www.nimh.nih.gov/health/statistics/major-depression.

Matteo module 4. "An Introduction to Natural Language Processing (NLP)." An Introduction to Natural Language Processing (NLP): 2.3 Word Count, https://port.sas.ac.uk/mod/book/view.php?id=583&chapterid=381

Ramirez-Esparza, Nairan, et al. "The Psychology of Word Use in Depression Forums in English and in Spanish: Testing Two Text Analytic Approaches." Proceedings of the International AAAI Conference on Web and Social Media, 25 Sept. 2021, https://ojs.aaai.org/index.php/ICWSM/article/view/18623.

Gandhi, Rohith. "Naive Bayes Classifier." Medium, Towards Data Science, 17 May 2018, https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c.

Stecanella, Bruno. "Understanding TF-ID: A Simple Introduction." MonkeyLearn Blog, 10 May 2019, https://monkeylearn.com/blog/what-is-tf-idf/.

Seldon. "Decision Trees in Machine Learning Explained." Seldon, 13 Nov. 2021, https://www.seldon.io/decision-trees-in-machine-learning#:~:text=Decision%20trees%20are%20an%20approach,categorise%20or%20classify%20an%20object.

Baumgartner, Jason, et al. "The Pushshift Reddit Dataset." ArXiv.org, 23 Jan. 2020, https://arxiv.org/abs/2001.08435.

Weissman, Benjamin. "Emoji Semantics/Pragmatics: Investigating Commitment and Lying." ACL Anthology, https://aclanthology.org/2022.emoji-1.3/.

Orabi, Ahmed Husseini, et al. "Deep Learning for Depression Detection of Twitter Users." ACL Anthology, https://aclanthology.org/W18-0609/.

Shing, Han-Chin, et al. "Expert, Crowdsourced, and Machine Assessment of Suicide Risk via Online Postings." ACL Anthology, https://aclanthology.org/W18-0603/.

Yla R. Tausczik, and James W. Pennebaker. "The Psychological Meaning of Words: LIWC and ... - Sage Journals." SAGE JOURNALS, https://journals.sagepub.com/doi/abs/10.1177/0261927X09351676.