Investigating Links Between Schizophrenia and Suicidal Language on Reddit

Jesenia Parthasarathy

Jesenia Parthasarathy

Introduction

People suffering from schizophrenia comprise only one percent of the US population, yet are four times more likely to die of suicide than the general US population.[1] This prompts the question of why this is true, and if it is possible to detect such thoughts before action is taken. In this data-driven study, I was loosely guided by the question of "what are the links betweens schizophrenic diagnoses and suicidal langauge on social media?", hypothesizing a higher correlation between schizophrenia and suicidal text. Multiple studies have investigated social media language as a tool for identifying mental health issues. I chose to closely follow Hswen, Naslund, Brownstein, and Hawkin's study monitoring tweets across self-identifying schizophrenic patients and determining the level of suicidal intent in their Twitter posts,[2] but additionally referenced methods from Coppersmith, Dredze, and Harman's study on quantifying mental health signals from Twitter posts.[3] My study collects 20 posts each from 148 unique Reddits users self-identifying as either schizophrenic, solely mentally unhealthy (and not schizophrenic), and neither schizophrenic or mentally unhealthy as a dataset, then uses a training dataset to train a machine learning model on what suicidal language is, then tests this model on the rest of the dataset to identify which has suicidal language.

This study can easily be connected back to themes in Big Data due to the very nature of its subjects. I collected data on clinical, or abnormal, psychology and investigated the mind, brain and behavior. Schizophrenia is a disorder defined by disruptions in mental processes, world perception, and social interactions.[4] In the data collection process, it was important to note the benefits and pitfalls of subsequent data-driven analysis. Regarding benefits, the data will often be quite naturalistic. Each text is from unsuspecting individuals who are unlikely to have a response bias, or to choose to respond differently than how they feel based on the study (i.e. people wishing to avoid suicide intervention might choose to not speak about suicide if prompted, despite truly feeling suicidal). The passive

[1] McManus, K., Mallory, E. K., Goldfeder, R. L., Haynes, W. A., & Tatum, J. D. (2015). Mining Twitter Data to Improve Detection of Schizophrenia. AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science, 2015, 122–126.
[2] Hswen, Y., Naslund, J. A., Brownstein, J. S., & Hawkins, J. B. (2018). Monitoring Online Discussions About Suicide Among Twitter Users With Schizophrenia: Exploratory Study. JMIR mental health, 5(4), e11483. https://doi.org/10.2196/11483
[3] Coppersmith, G., Dredze, M., & Harman, C. (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
[4] U.S. Department of Health and Human Services. (n.d.). Schizophrenia. National Institute of Mental Health. https://www.nimh.nih.gov/health/statistics/schizophrenia#part_2546

nature of collecting data from the internet avoids intentional skews of responses. Additionally, the all_posts dataset is large enough to be able to make future predictions and to use as a training set for testing responses in the future, as well as find steady trends and patterns that become clearer with larger study sizes. In regards to cons, it is imperative that a researcher watches out for data-fishing, or the practice of forcefully producing a significant result without correcting for comparisons. The final prediction dataset (final_predict) consisted of 2318 rows (with 148 users/participants) and 4 columns (Username, post text, diagnosis of either schizophrenic, mental health, or neither, and prediction of suicidal text), and the final post dataset consisted of 2316 rows (same participants) and 3 columns, excluding the prediction column. These are both examples of wide data, with many subjects across a wide-range of new individuals but not very much data on each point.

Methodology

I used a data-scraper with Python, namely the Reddit API. I initially had hoped to use Twitter, however, the Twitter API required payment plans to use it for the purposes I had hoped to be able to use it for (i.e. scraping users and their recent 20 posts). Therefore, I opted for Reddit data, which might even be better suited since text limits are less stringent than Twitter's. On top of cost limitations, APIs often have access limitations, through access rates, however for the data I was trying to collect, I had no trouble abiding by the rule of 100 queries per minute.[5] With Reddit's API, the python command to call this API requires the in-built tool called PRAW (Python Reddit API Wrapper), a convenient way to call Reddit data straight to my terminal. This way, I can avoid storage space issues with big data by not transforming my data into a csv file, and opting to keep all the data on the browser kernel. However, due to requirements to submit data, I downloaded the data regardless.

To begin, I had to register an account on Reddit and collect credentials so that Reddit knows each request is connected to my account. This included a "client ID", "client secret", and "user agent." I chose to scrape the subreddit r/schizophrenia since this is a popular subreddit, with 79,000 members

[5] Reddit API Documentation. https://www.reddit.com/dev/api/

and in the top 2% of popular subreddits,[6] where those suffering from schizophrenia often post their experiences. I manually scrolled through subreddit to find 50 unique usernames, where each self-identifies in either a 20-most recent post or in their bio that they are schizophrenia, schizoaffective, schizotypal, and/or have psychosis, checking for bots and only choosing those with at least 20 posts. This has limitations: (1) the process of manually picking 50 usernames took 40 minutes, (2) each diagnosis is self-identified, making it unclear if each diagnosis is medically accurate, and (3) there is a selection bias towards people who have posted at least 20 times. In a reference study from De Choudhury, Gamon, Counts, and Horvitz, the researchers chose to personally interview medically-diagnosed schizophrenic patients, and then collect their Twitter users.[7] This is a method to correct against naive assumptions of honesty.

After saving this information into a dataframe, I repeated this same process with subreddit r/mentalhealth,[8] of 463,000 members, and manually identified 50 usernames that identify across a range of mental health issues (i.e. depression, anxiety, ADHD, bipolar, OCD, etc.). These might double-identify, as such some of these users might also be schizophrenic, however since I cannot control for this, the only way to ensure they do not overlap is by choosing users who mention mental health issues and do not mention schizophrenic diagnoses. Additionally, those identified in the schizophrenic category might also be mentally unhealthy, however since this is a lot harder to control for (since psychiatric comorbidities are common among patients with schizophrenia),[9] it was not considered in text-collection. Finally, I needed to pull 50 random users on Reddit as a control group. However, this was nearly impossible since the API does not permit doing so, and pulling random users from a random subreddit will bias the results in favor of the type of people using that subreddit. Therefore, I used a random subreddit generator,[10] where in each I picked the most recent non-bot post having 20+ posts and saved the poster's username. I repeated this for 50 different random subreddits.

---

[6] Reddit subreddit. r/schizophrenia. https://www.reddit.com/r/schizophrenia/
[7] De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2021). Predicting Depression via Social Media. Proceedings of the International AAAI Conference on Web and Social Media, 7(1), 128-137. https://doi.org/10.1609/icwsm.v7i1.14432
[8] Reddit subreddit. r/mentalhealth. https://www.reddit.com/r/mentalhealth/
[9] Buckley, P. F., Miller, B. J., Lehrer, D. S., & Castle, D. J. (2009). Psychiatric comorbidities and schizophrenia. Schizophrenia bulletin, 35(2), 383–402. https://doi.org/10.1093/schbul/sbn135
[10] Perchance random subreddit generator. https://perchance.org/subreddits

Finally, I combined these 148 usernames (two were deleted during the span of this study from an original 150 usernames) and their 20 posts each into one dataframe. Limitations of this data collection method include not collecting for demographics, such as gender or age distribution. Hswen's identifies these demographics since age and gender have a strong impact on rates of suicidal langauge as well as the type of suicidal langauge.[11] Additionally, some posts were either blank or failed to be collected by the API (recorded as [removed]), so despite expecting 3000 rows of data, my final data set had 2315 rows.

After collecting this initial data, I had to collect a training set to identify suicidal language and therefore identify each post as either suicidal or not. I scraped the subreddit, r/SuicideWatch[12] and manually identified 19 suicidal texts, labeling each as "True"—a boolean. I utilized the random subreddit generator method to find 18 random posts that I personally identified to be non-suicidal language (labeled "False".) I edited this small dataset in Excel to ensure quality; it contained 37 rows with each row characterized by "compiled_text" (the post text) and an identification of true or false under "is_suicidal". I begin by training a Logistic Regression model with the scikit-learn package on Python, first vectorizing the text with max_features set at 1000, and then testing this model on my original text dataset of 2315 rows. However, upon completion of testing, we can clearly see that the model might not be 100% reliable at detecting suicidal language. To demonstrate this, I picked a random text piece:

"Can anyone of you tell me what this machine is and is doing that i rode past. They are building a residential building and to me it looks like maybe creating more foundation?"

This text is identified to be true (is suicidal). This prompts investigation of our limitations: First, the size of our training set violates best-practices of the "hold-out method" that encourages splitting the data where 80% is used to train and 20% is tested on. My method only trains on approximately 1.7% of the original data set size. To properly abide by this rule, I would need to get a dataset of approximately 1840 rows, with half being suicidal language and the other half being not. Myself, being the only

[11] Monitoring Online Discussions About Suicide Among Twitter Users With Schizophrenia: Exploratory Study
[12] Reddit subreddit. r/SuicideWatch. https://www.reddit.com/r/SuicideWatch/

researcher, and having a mental limit for how much suicidal text I can personally parse through, manually finding around 920 suicidal posts would be infeasible. The original study I followed had multiple researchers working for a longer time span to identify language in Twitter posts which is a method to offset this issue.[13] Second, despite vectorizing our text, Python has a Natural Language Toolkit (NPTK) with large corpi that can help identify sentiment and tokenize text language that I had not utilized. Therefore messy text could be a reason why I got occasional unreliable results. Third, the suicidal posts often have buffer-text such as "I looked at my house today," before mentioning actual suicidal text, and the model is trained to identify the former as undoubtedly suicidal text. This therefore means false positives are highly likely with my model. This can be fixed by instead using a self-made corpus of suicidal langague, similar to the NPTK toolkit's corpus of sentiment analysis. Fourth, and final, is that the machine learning algorithm I chose to utilize was Logistic Regression, whereas there can be more advanced utilizations of Support Vector Machines (SVM) which I was limited from due to technical skill issues. These advanced machine learning techniques would reduce the level of error my model ended up getting.

A final note regarding the coding. Initially, I identified 50 valid usernames from each group, totalling 150 usernames. However, I revisited the scraping code a week later and two of these usernames had actually been disabled (one from the mental health group, and another from the either group). This then broke my code, requiring me to add checks back against usernames that deleted themselves. I ended with 148 usernames instead.

Results

Notably, despite errors in predictions identified in the previous section, we can likely still search for trends due to the large nature of this dataset. First, we can take a look at the proportion of True (suicidal) vs. False (non-suicidal) predictions by the model by diagnosis. For those with neither diagnosis, only 15% of the total posts were identified by the model to be suicidal. There is a possible issue here, based in the lack of medical accuracy, in that the people who do not discuss mental health

[13] Monitoring Online Discussions About Suicide Among Twitter Users With Schizophrenia: Exploratory Study

issues on social media, are likely also not going to discuss suicidal thoughts on social media. This is still a significant conclusion, since the proportion of people not discussing their diagnosis that might actually still discuss suicidal intentions is 15% (which is quite high for a group of people who do not identify as mental unhealthy on these platforms). On the other hand, nearly 46% of the posts from self-identifying schizophrenics were identified as suicidal and a close 42% in the mental health group. It can be noted that there might be a confounder, or mis-training of the model, in that many of the suicidal training set posts mention some issues regarding mental health, and are less likely to mention schizophrenia. As a result, the model begins to identify, generic text about mental health issues, as suicidal text. This might exacerbate the rate of false positives in the mental health group. When looking at the users on their own, the percent of all mental health users that have at least one suicidal text is around 96% and for schizophrenia, around 90%. There is a clear dropoff with those in the neither category, as they have only 53% of users with at least one post identified as suicidal. Again, this can be seen as a significant deviation of expectations, despite not being as high as the percent of users in the other categories.

The original study checked directly for words like "suicide" and "suicidal" to determine finding suicidal language.[14] The issue with this, is that text might refer to suicidal intentions without ever actually explicitly saying "suicide" and sometimes, text might say suicide in other contexts aside from the feelings of the poster. For example, I compiled all the users that have posts that mention "suicide" or "suicidal," and decided to investigate one of them. Out of the 13 users, each having one post explicitly referencing suicide, five had a mental health diagnosis, seven had a schizophrenic diagnosis, and only one had diagnosis of neither. I pulled up the post from the neither diagnosis, and found my model had not predicted any of their posts to be suicidal language. After further investigation, I found that their post had been the link to an article about suicide in Iraq. This shows a pitfall, in which had I tried to replicate the original study without thorough investigation, I would not have been able to recognize these outliers the way they did.

[14] Monitoring Online Discussions About Suicide Among Twitter Users With Schizophrenia: Exploratory Study

Finally, I did a Chi-Squared Test and found a chi2 statistic and p-value regarding the relationship diagnosis and the prediction for suicidal text. The chi2 statistic is a measure of how much observed frequencies deviate from the expected frequencies, but more importantly, the p-value is the probability that the observed association is due to chance. I chose 0.05 as an alpha level, or significance level, and if the p-value was below that alpha level, I would reject the null hypothesis of there not being a relationship (due to chance). The p-value was 2.018×10^{-18}, signifying a likely association between schizophrenia diagnosis and suicidal text, rejecting the null hypothesis. I extended this by looking for the Cramér's V and R-squared across the three diagnoses and the prediction of suicidal language. I had to first encode the diagnoses through hot-one encoding. This creates three columns, and allows for booleans in each row to identify the true diagnosis (i.e. if the diagnosis for a user is mental health (MH), then there would be a "true" in column Diagnosis_MH, a "false" in Diagnosis_SCH and "false" in Diagnosis_Neither). After encoding, I could finally find the actual test values. The Cramér's V ranges from 0 to 1, where 0 is no association and 1 is perfect association. In this test, the value was 0.1819, which indicates nearly a weak to moderate association. Additionally, I found the R-squared value, a value ranging from 0 to 1 where 0 indicates that the independent variables do not explain any of the variance in the dependent variable and 1 indicates full explanation of variance. In this model, however, it seems that only 8.46% of the variance suicidal predictions are explained by the diagnosis.

In conclusion, initial investigation finds differences in the type of language, however after checking with statistical testing, we find less significance in these relationships. Therefore, it can be helpful to expand these databases and refine the prediction model, then re-perform these statistical tests to see if they match results from initial investigations.

Discussion

The original question guiding data-driven analysis was "what are the links betweens schizophrenic diagnoses and suicidal langauge on social media?"—and a specific interest in seeing a higher degree of suicidal text in schizophrenic populations on social media. We found a strong correlation between the diagnosis of schizophrenia and being predicted as suicidal, however due to the

model's instability, it is hard to explain the variance in suicide predictions to the diagnosis alone, i.e. there are definitely other factors in the text that determine whether the model is able to predict it as suicidal. After adjusting for issues in the model's building, these can be regulated better. Additionally, there are a few implications from both the data collection and the model for the real world.

First regarding the data collection: it is important to note the ethical implications of scraping personal words from what can essentially be considered people's diaries. Despite this being good for reducing response bias, there are inherent questions on how much personal information we can compile for a given user. To avoid these issues, I avoided collecting demographic data, however there are still questions as to whether even identifying the username in the first place is ethical. Due to these implications with demographic data, any usernames in the original study were removed, whereas, since my dataset has only publicly accessible information such as username and their posts, I will opt to retain these usernames. The second implication is what intervention opportunities can arise from training machine learning models to be able to identify suicidal text. Even now, searching up suicidal language in Google's search engine triggers intervention methods such as printing the suicide hotline's phone number. Similarly, the third implication is how peers can identify suicidal language within texts. Additional information regarding early hints, or demographic information such as psychiatric comorbidities, etc., are all valuable information that help train regression models to identify suicidal language, and can similarly be used to train humans to identify this language.

Regarding the prediction model: After adjusting for the limitations clearly highlighted in the methodology portion of this paper, this model can begin to do the reverse, i.e. predict mental health issues or schizophrenia. Intervention becomes easier after the monitoring of mental health or psychiatric issues are identified. This is a lot more effective since the training set would essentially just be this whole dataset of 2000+ rows. The hindrances I faced collecting this data demonstrates that in this specific method of data collection, due to the requirement to have humans checking against the machine, it helps to have multiple researchers on the project in order to expand its size, not just with collection but also with cleaning.