Effects Of Social Media On Mental Health

Neel Mehta
nmehta32@uic.edu
University Of Illinois at Chicago
Chicago, Illinois, USA

Sai Lahari Sanku
ssanku4@uic.edu
University Of Illinois at Chicago
Chicago, Illinois, USA

Shreya Raj Kati
skati6@uic.edu
University Of Illinois at Chicago
Chicago, Illinois, USA

Mohammed Tanveer
mlnu23@uic.edu
University Of Illinois at Chicago
Chicago, Illinois, USA

1 ABSTRACT

Depression is one of the most common mental illnesses in the country, especially for young adults. Nearly one-third of university students struggle with depression and other forms of mental health disorders. This is slowly becoming an epidemic in the western world. With the increase in the usage of social media, studies associate the two with each other whether there is a causality between both remains to be seen. We on the other hand dive deeper into the usage of social media and try to better comprehend the causal relationships between support received on social media posts (Tweets) and the re-occurrence of depressive posts. We collected data in the form of tweets, cleaned the data, performed exploratory data analysis, ran a machine learning algorithm to classify depressive tweets, and then calculated the average treatment effect of the intervention based on the support a depressive tweet got in the form of favorites, retweets, and quotes.

2 INTRODUCTION

Mental health illnesses are one of the greatest challenges that our society faces. More than one-third of the younger population undergo depression, which has increased especially since the pandemic began. Most of the population that has symptoms of depression are not treated adequately, especially with United States healthcare not being affordable to the vast majority of the population. Furthermore, ethnic minority groups such as Mexican Americans and African Americans are significantly less likely to receive depression treatments than other ethnic groups. This places a huge societal problem, especially in fast-paced, ever growing environments in metropolitan cities and first-world countries like the United States, as it not only weakens the sufferer, but has a direct impact on their physical health, as depression increases the chances of getting chronic illnesses, such as cardiovascular disease, diabetes and obesity to name a few. Additionally, it is more likely that someone who is struggling with depression will experience negative effects on their daily activities, such as being unable to maintain a regular sleep schedule and picking up bad eating habits, as well as their relationships with their family, friends, and coworkers.

Social media was created to bring people together, especially when they are lonely or depressed. We have more social media apps and connectivity than ever before, so when people are depressed, they can express their frustrations and thoughts to the public in order to seek help or be heard by their friends. The widespread use of social media opens up numerous possibilities for analyzing the user base's mental well being. Twitter and other social media platforms could be useful tools for research on mental health because they provide a unique, measurable perspective on human behavior that might otherwise go unnoticed. Because of Twitter's relatively anonymous communication platform, a broader range of people who might not normally participate in research can provide a less biased, realistic depiction of people's experiences. Twitter, despite being accused of using stigmatizing and discriminatory terminology, is a valuable and less intrusive way to learn about the discourse surrounding mental health from both those experiencing mental illness and the general public. The information in tweets about depression is still relatively new. Depression symptoms have been noted in Twitter posts, but it is unknown if the extent of supportive actions over tweets (i.e. likes, retweets, and the number of followers) have any effect on depression, although it is entirely possible that tweeters actively seek out or post supportive comments or advice about depression on Twitter, and in the other end, people on twitter actively seek out to humiliate and/ or make their friends or even strangers feel worse on tweets that are depressive.

Using social media as a tool for behavioral health assessment has advantages because, unlike self-report methodology in behavioral surveys, where responses are incited by the experimenter and typically consist of remembering of (occasionally interpretive) health facts, social media measurement of behavior captures social activity and language expression in a naturalistic setting. Such activity occurs in real time and in the course of a person's daily life. As a result, it is less susceptible to memory bias or experimenter demand effects, and it can assist in tracking concerns on a fine-grained temporal scale.

Using this as our motivation we were able to achieve the following:

Using Twitter scraping techniques, a dataset of 23,791 tweets from 870 different users was collected. This was accomplished with the Twinit was used to scrape user data from Twitter, and tweepy was used to scrape individual tweets. We used a statistical model (a transformer-based model classifier) to predict whether the message in the Twitter post is depressive or not. To classify the text, we used the deep BERT (Bidirectional Encoder Representations from Transformers) model. We proposed two new metrics, 'Support' and 'Score,' with 'Support' based on the tweet's engagement and 'Score' based on the difference between the percentage of depressive tweets from an individual user after the intervention and the percentage of depressive tweets from an individual user before the intervention. We used the variables to estimate the outcome effect and the Average Treatment Effect (ATE) and used various matching

methods such as Propensity Score Matching, Mahalanobis Distance Matching, and Nearest Neighbor Matching.

After completing the aforementioned tasks, we believe we gained a better understanding of the causal relationships between users and their behavior on the social media platforms Twitter.

3 RELATED WORK

There is evidence[1][2] showing that Social media usage among people with a variety of mental problems, such as depression, psychotic disorders, or other serious mental illnesses, is equivalent to that of the general public. As a result of this, there has been a lot of development recently on using social media data to obtain meaningful insights and inferences. A lot of these approaches use Natural Language Processing techniques like n-grams to obtain evidence of mental illness based on the language used[3][4]. The framework proposed in [5] provides a good way to identify depression based on the language used on social media forums.

In the context of causal inference, The work done by the authors in [6] provides estimation and inference for causal effects on a single social network by using a structural equation model. The important aspect of this paper is that it included dependence due to the transmission of information across network ties in their estimation. Authors in [7] used blogs to forecast the country's suicide rate statistics. Authors in [8] surveyed a sample of Twitter users to investigate the relationship between tweets about suicide and suicidal behavior while researching the linguistic characteristics of suicidal thoughts. If supporting people would cause them to tweet about topics that would be considered depressive, that would be a good chunk of progress made as many studies show that depression is decreased when you have peer support [9].

However, the majority of this earlier research explored distinctions between content connected to suicide and other content posted on social media or focused on macro-level trends (such as national suicide rates).

The estimate of treatment effects in observational research is fundamentally more difficult compared to doing the same in well-designed experiments because of selection bias. Authors in [13] discuss how to match data using the Mahalanobis distance to lessen covariate imbalance and increase the accuracy of treatment effect estimations. The authors in [11] provide the idea of intervention studies and the use of Mahalanobis' distance for a health promotion research control group selection. In [14], propensity score-based methods were used for the analysis of observational data. The authors consider propensity score as a balancing score, that helps maintain the distribution of observed baseline covariates will be similar between treated and untreated subjects. In [15], the authors use propensity scores to reduce the effects of confounding when estimating the effects of treatments.

4 FORMAL PROBLEM DESCRIPTION

Our problem description states whether interaction metrics like the number of replies, favorites, quoted tweets, and retweets affect the number of tweets indicating depression by a user after the given intervention. We assume the interaction with the tweet of the users, such as likes, replies, quotes, and retweets of the original tweet, to be a form of "support". Through this project, we main to determine if the "support" has an effect on the number of tweets that are depressive in nature by a user after an intervention.

The primary goal of our research is: To determine the impact of support on a depressive tweet in the form of engagement on the difference in the number of depressive tweets 'pre-support' and 'post-support'.

We gathered and analyzed the scraped dataset of depressive-related and 'normal' tweets to accomplish this. Our main task is to estimate the causal effect of receiving 'high' support on a depressive tweet and the user's behavior following high 'support'.

5 PROPOSED SOLUTIONS

Our approach follows the framework mentioned in figure 2. The process is explained below - First, in order to get labels for whether the posted tweet is depressive or not, we performed term-frequency inverse document frequency (TF-IDF), where it's used to calculate how relevant a term is in the given document and not just having a higher frequency in the document.

The inverse document frequency (IDF) of a word across the data set suggests how rare a word is in the entire document; if it's closer to 0, the word is more common, and 1 otherwise. This metric can be calculated by taking the total number of tweets in our case, dividing it by the number of tweets that contain a word, and calculating the logarithm. Multiplying term frequency with IDF gives us TF-IDF.

$$\text{TF(i,j)=n(i,j)}/\sum n(i,j)$$

$$IDF = 1 + log(N/dN)$$

$$TF - IDF = TF * IDF$$

We trained the classifier using a public Twitter data set that contained only the posted tweets and their respective "depressive or not depressive" labels, The data set had 20,000 tweets and corresponding labels. We split the data set into a 70-30 split, which gave us 14,005 tweets for the training data and 5,995 tweets for the testing data. Training the classifier on the training dataset and then running the classifier on the test dataset, we got the following metrics: Precision: 0.7959839357429719, Recall: 1.0 Accuracy: 0.8974151857835219 and F-score: 0.88. The dataset used for training is different from our original data set on which we calculate the treatment effect. With this classifier, we then classified the tweets in our main dataset and marked '0' for the non-depressive tweet and '1' for the depressive tweet. After data pre processing, we used BERT to classify our tweets, in order to achieve this we used the 12 encoder with 12 bidirectional self-attention heads, this is also called BERT_{BASE}.

We went ahead with this as our classifier because BERT takes into account the context for each occurrence of a given word instead of just frequency and importance of the words used by the previous context-free models such as word2vec, and TF-IDF. As the previous models considered the word 'running' to have the same meaning (embedding), "He is running a company" and "He is running a marathon". But as we know that, the use of the word 'running' here has different meanings. Hence we went ahead with using BERT as our model to classify the tweets.

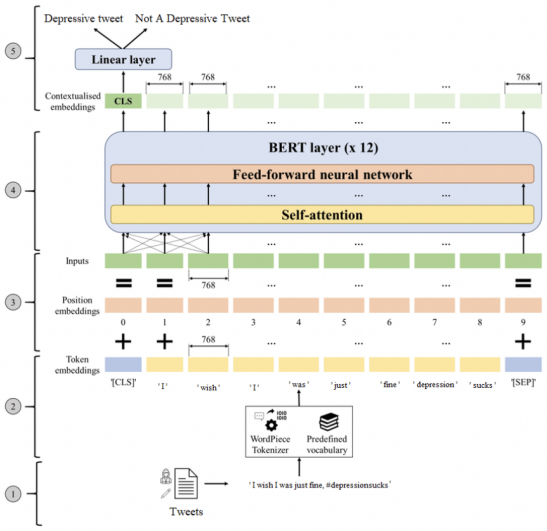

Figure 1: The tweet as a sentence is first preprocessed and sent to the BERT model. Where the sentence is tokenized and each token is given a position embedding which makes up the 768 input embeddings to the BERT layer. The BERT layer gives us 768 contextualized embeddings as output, and when passed through a linear classification layer, we get the label for the tweet.

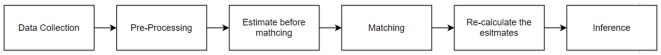

Figure 2: The framework of the proposed solution

We trained the model with a training dataset consisting of 20,000 tweets which had a flag variable to indicate whether the text was depressive or not. This dataset was then split into 70-30 split of training and test data and then we ran the model on the training data. After the training was completed on all the training tweets, we stop running the model on the training data. We then had the embedding for the network, and then we ran the model on the test data and got an accuracy of 98 percent. We saved the model and then we ran it on our original dataset, and acquired the labels to the tweets.

The 'Support Level' was calculated using the 'Support' which has the following formula:

$$\text{Support} = (\text{No. of Likes} + \text{No. of Replies} + \text{No. of Quotes} + 1) / (\text{No. of RT's} + 2)$$

We decided to normalize it with the number of retweets as the denominator because we did not have the number of followers of the user, and from prior domain knowledge, we know that generally

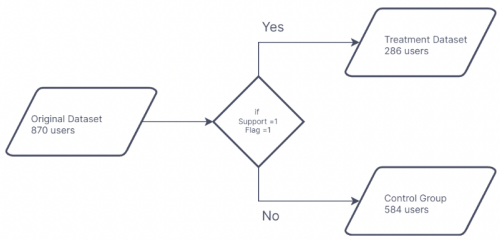

Figure 3: Splitting the dataset into control and treatment groups

only people with a high amount of followings get a high number of retweets, as compared to other users.

Furthermore we had to split the users in the groups of control and treatment. We were able to achieve this by randomizing a tweet and assigning the user of that tweet into control/ treatment based on the 'Support Level' and the Depressive 'Flag' variable. If the tweet had 'Support Level' = 1 which indicates as high, and if the depressive Flag was '1' indicating the tweet was depressive, then the user is assigned to the treatment group, otherwise the user is assigned to the control group. We were able to get 286 users in the treatment group and 584 users in the control group, as seen in the figure 3.

After assigning the users to the control group we then calculate the outcome, which is the difference between the percentage of the number of depressive tweets post-intervention and pre-intervention.

$$\text{Score} = (\text{Number of depressive tweets after the intervention}/\text{Total number of tweets after the intervention} - \text{Number of depressive tweets before the intervention}/\text{Total number of tweets before the intervention}) * 100$$

After calculating the score for each user we then average it for the control group and treatment group, this gives us the average treatment effect on control and the average treatment effect on treatment. And Average Treatment Effect is calculated by taking the average on the whole population i.e. the users. But this was done before we even performed any matching, we need to perform matching to find pairs, and reduce model dependence and reduce bias due to confounding.

We use several matching techniques such as the well known propensity score matching (PSM), Nearest Neighbor Matching, and Mahalanobis Distance Matching. In some of the methods we also use linear regression to fit the data after matching. Then we calculate the estimates using the library CausalModel in python, then we are then able to provide inference based on the values of the estimate we get.

5.1 Propensity Score Matching (PSM):

Propensity score matching (PSM) is a statistical matching technique used in the statistical analysis of observational data that aims to

estimate the impact of a treatment, policy, or other intervention by taking into account the covariates that determine whether a person will receive the intervention. PSM makes an effort to lessen the bias brought on by confounding factors that could be present in an estimate of the treatment effect discovered by only contrasting the outcomes between units that got the treatment and those that did not.

In normal matching, single characteristics that distinguish treatment and control groups are matched in an attempt to make the groups more alike. But if the two groups do not have substantial overlap, then substantial error may be introduced. For example, if only the worst cases from the untreated "comparison" group are compared to only the best cases from the treatment group, the result may be regression toward the mean, which may make the comparison group look better or worse than reality.

We had the features like length of tweet, support the user recieved and the user characteristics like number likes replies etc. We ran this through a logistic regression model to get the propensity score of every user in the treatment and control group. The users in the control group were than “matched” to the users in the treatment group by exact matching ie users with the exact same propensity score were matched together users that were not matched were removed from the dataset. This was done because of the high level of propensity matching that was obtained for treatment and control groups (figure 8). After Matching we obtained 200 users each in the treatment and control group (figure 9).

After PSM we get a new cleaner database with less model dependance. Hence, we get more pronounced results:

ate before matching 0.37648642041469493 ate after matching 1.1206713795994439

This supports our original result and hypothesis that if support is presented to users on a depressive tweets they are less likely to tweet depressive things in the future.

5.2 Mahalanobis Distance Matching

The Mahalanobis distance[10] is the distance between two points in a multivariate space. It resembles the Euclidean distance in certain ways. The main distinction is that while Euclidean distance uses the original data, Mahalanobis Distance Matching (MDM) uses the standardized data[11].The distance between two points is calculated using the formula[12] -

$$\text{Distance(Xc, Xt)} = \sqrt{(Xc - Xt)'S - 1(Xc - Xt)}$$

Where Xc belongs to Control group and Xt belongs to treatment group. Based on this distance, every unit in the treatment group is mapped to a unit in the control group. Units that could not be mapped will be pruned.In addition to this, we also use a caliper, which is the acceptable maximum distance threshold. If the distance between any pain is greater than the caliper, then that pair is pruned. Using this approach, we aim to estimate the difference in means of the treatment effect of both the groups, thereby estimating causal effect.

6 DATA DESCRIPTION AND CLEANING

Twitter is a popular micro-blogging service where users create status messages called tweets. These tweets sometimes express opinions about different topics. With the help of the Twitter API, it is easy to extract large amounts of tweets with emoticons.

We used Twint, an advanced Twitter scraping tool in Python that allows for scraping Tweets from Twitter profiles without using Twitter’s API. We also used Tweepy, an open-source Python package that gives you a very convenient way to access the Twitter API with Python. Tweepy includes a set of classes and methods that represent Twitter’s models and API endpoints, and it transparently handles various implementation details, such as Data encoding and decoding. We used Twint to extract the user-id for the users diagnosed with depression. Once, the user ID was found we used Tweepy to extract the tweets by the particular users.

We collected these tweets by first getting user accounts that had tweeted the phrase "diagnosed with depression", every user in our dataset has self identified that he/she is diagnosed with depression. As a result, we have collected novel twitter data which includes 23791 tweets from 870 Twitter users. (tanveer please add how we got treatment and support groups)

They contain the following labels:

- **Date & Time:** The date and time in UTC timezone when this Tweet was posted.
- **Tweet Id:** The unique identifier for this Tweet.
- **Text:** The actual tweet text of the status update.
- **Username:** The user who posted this Tweet.
- **Likes:** The favorite count indicates approximately how many times Twitter users have liked this Tweet.
- **Quote Count:** The Number of quotes of the tweet, Indicates how many times Twitter users have quoted this Tweet.
- **Reply Count:** Number of times this Tweet was replied to.
- **Retweet Count:** Number of times this Tweet was retweeted.
- **Android Or Apple:** The device used for the tweet.

In addition to these variables we decided to add 4 more variables to the dataset, those are as follows:

- **Tweet Length:** The number of characters in the tweet:
- **Time of Day:** The time at which the tweet was posted. (1: ‘Morning’ to 4:’Night’)
- **Support Level:** This indicated the support the tweet got. (1: ‘High Support and 0: Low Support)
- **Flag:** This indicated whether the tweet was flagged depressive or not. (1: ‘Depressive’ and 0:’Non Depressive’)

In order to get some of these new variables we needed to clean the data. In order for us to pass the tweet text to the classifier, we needed to clean the text, so we decided to remove all punctuations, remove the tags of other users (such as ‘@usernamewhenreplyingto’), emoticons, and other symbols as well. This was done so that we get better results, when we pass the tweet text to the classifier that will better indicate whether the tweet based on the text was depressive or non depressive.

6.1 Exploratory Data Analysis & Classification

About the data it reflects, a tweet is full of views. Raw tweets are very unstructured and include redundant information when they are not first processed. Pre-processing of tweets is done in several

Figure 4: Word cloud generated from all the tweets in the dataset

ways to address these problems. The hashtags, emoticons, and slang terms used on nearly all social media platforms are associated with the subject matter they represent, which has an impact on data analysis and classification. NLTK (Natural Language Toolkit) is one of the best libraries for pre-processing text data. By using NLTK, we were able to tokenize, remove stop words, and then perform lemmatization to obtain cleaned words for our classification. Figure 1 shows the word cloud generated from the cleaned data from tweets. Also using NLTK library we were able analyse and obtain top frequent unigrams, bigrams and trigrams as shown in figure 2.

Figure 3 gives a list of the high-frequency unigrams that appear in the postings of the two classes. For the negative class (standard posts), most of the unigrams relate to commonplace details of daily life, ranging from work to entertainment (good, love, like, would, want, think, people). Some positive emotion words are also observed, such as love, like, and good. On the other hand, in the case of depression-indicative posts, many words are emotional (e.g., depression, diagnosed, mental, anxiety, help, disorder). The strong propensity for unigrams that express negative affect and low-intensity emotions, however, may be a reflection of the people who shared these messages' mental instability and sense of powerlessness, as well as signs of their likely depression. Additionally, some blogs make mention of medication and therapy, probably because the authors want to share information about these topics with their readers.

After performing data cleaning and exploratory data analysis we have applied classification algorithm on our tweet text data to classify where the tweet is depressive or not depressive. We used BERT classification algorithm for our study.

BERT is an acronym for Bidirectional Encoder Representations from Transformers, and this was preferred by us in the end because BERT does not only use the frequency of the words to classify, but also looks for the context the word was used. In many natural language problems, BERT outperforms the state-of-the-art by assisting machines in learning excellent representations of text in relation to context.

In order to train the classifier, we used a labeled data set that had labels for tweets being depressive, this was our training set. This had around 20,000 tweets with 10,000 having the label '1' to be depressive and 10,000 tweets having the label '0' to denote not depressive. For the classifier, we were able to achieve an Accuracy of 0.89 and an F-score of 0.88. This could be improved as we could

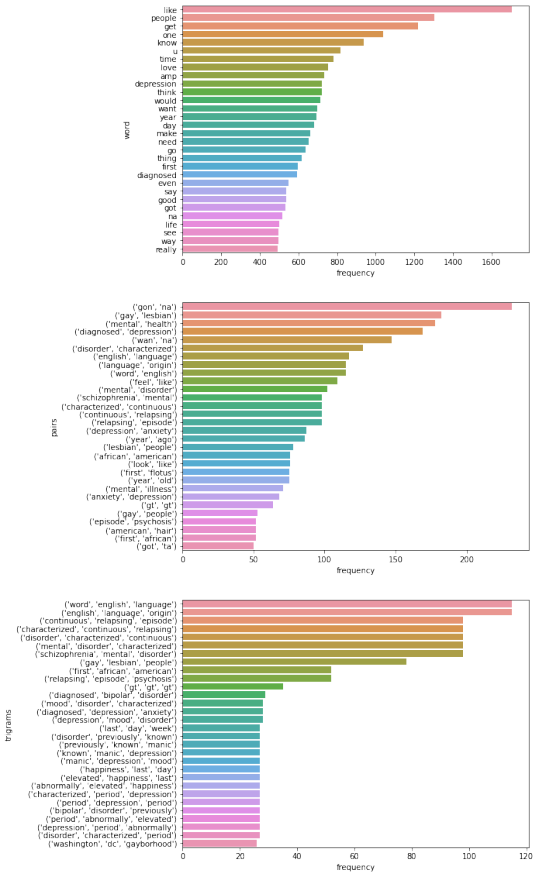

Figure 5: Bar graphs of frequent unigrams, bigrams and trigrams in the dataset

DEPRESSION CLASSIFIED TWEET WORDS	OTHER STANDARD TWEET WORDS
depression, diagnosed, like, people, get, mental, know, anxiety, one, year, disorder, time, health, life, think, would, day, u, need, thing, want, way, help, even	good, love, like, would, want, think, people, make, first, day, thing, time, say, one, need, year, know, got, go, get, gay, even

Figure 6: Top Unigrams for the two classes of tweets

identify the classifier mislabelling the tweet when the tweets contained only one word, and it mislabelled tweets containing slang such as 'kms' which is slang for 'kill myself'. The reason for the improvement is that our scrapped data is unlabeled so there is no sure way to get the accuracy. The classifier is trained on a different Twitter data set entirely so the actual accuracy of the classifier is lower than 0.9 on our scrapped data.

Figure 4 shows the distribution of users across both the classes after applying our BERT Classifier. Out of 870 total users in our

TOTAL NO OF USERS	**870**
(1) # USERS DETECTED WITH DEPRESSION	215
(2) # USERS WITH NO SIGNS OF DEPRESSION	655
TOTAL NO OF TWITTER POSTS	**23791**
(1) POSITIVE CLASS (DEPRESSIVE INDICATION)	4717
(2) NEGATIVE CLASS (STANDARD TWEETS)	19074

Figure 7: Statistics of Twitter data after classification

study, we were able to classify 215 users with depression indicative tweets and other 655 users with no or little signs of depression. Also we have classified 4717 tweets to be depressive in nature and other 19074 tweets as not depressive.

The major trends in our depressed tweets data set can be seen from the word cloud and bar chart shown. We were able to see people that who were depressed were more likely to tweet about depression, drugs, and curses a bit more than their counterparts. It was interesting to see drugs being pretty frequent in the data set and upon further inspection of the tweets, we found it to make sense as most people are inadequately treated, so depressive people are more likely to turn to illicit drugs to cope with their depression.

7 EXPERIMENTAL SETUP AND RESULTS

Initially, we scrapped the data from Twitter for the search tag " diagnosed with depression" to find the users who are depressed. And obtained about 300 users who are depressed. And then obtained the other 50 tweets for each of these depressive users. Similarly, 300 users in the control group were selected limiting the number of tweets per person to 50. We then performed exploratory data analysis to plot the word cloud for depressive users (figure 1) and a bar graph to visualize the distribution of words across our data set (figure).

We select a random depressive tweet per user, ie, a tweet with the label 1. We calculated the label "Score" by (ReplyCount + RetweetCount +1 / LikesCount +2). We used Laplace smoothing to make the data consistent and to make sure that we don't divide by 0. This score was calculated for 2 instances. For the tweets before the intervention (random depressive tweets) and for the tweets after the intervention. Here this gives us our Treatment effect, which can be inferred by how much effect of support is received by the user who posted about their diagnosis with depression. After obtaining the treatment effects for all our 282 users we calculated the Average treatment effect.

We have achieved about 1.96 of the average treatment effect of the intervention ranging between 0.05 to 16.5 as shown in figure 3.

7.1 Propensity Score Matching

Propensity score matching (PSM) is a statistical matching technique used in the statistical analysis of observational data that aims to estimate the impact of a treatment, policy, or other intervention by taking into account the covariates that determine whether a person will receive the intervention. PSM makes an effort to lessen the bias brought on by confounding factors that could be present in an estimate of the treatment effect discovered by only contrasting the outcomes between units that got the treatment and those that did not.

In normal matching, single characteristics that distinguish treatment and control groups are matched in an attempt to make the groups more alike. But if the two groups do not have substantial overlap, then substantial error may be introduced. For example, if only the worst cases from the untreated "comparison" group are compared to only the best cases from the treatment group, the result may be regression toward the mean, which may make the comparison group look better or worse than reality.

We had the features like length of tweet, support the user received and the user characteristics like number likes replies etc. We ran this through a logistic regression model to get the propensity score of every user in the treatment and control group. The users in the control group were than "matched" to the users in the treatment group by exact matching ie users with the exact same propensity score were matched together users that were not matched were removed from the dataset. This was done because of the high level of propensity matching that was obtained for treatment and control groups (figure 8). After Matching we obtained 200 users each in the treatment and control group (figure 9).

After PSM we get a new cleaner database with less model dependence.

Hence, we get more pronounced results:

ate before matching 0.37648642041469493 ate after matching 1.1206713795994439

This supports our original result and hypothesis that if support is presented to users on a depressive tweets they are less likely to tweet depressive things in the future.

7.2 Mahalanobis Distance Matching

After assigning users to control and treatment groups, based on the support and the label indicating depression, we compute the average treatment effect of each group and compute the difference between the means. The difference before matching is described in figure 10. After this, we perform mahalanobis matching on the data. The caliper value for this experiment is 0.2. This seemed to be the best value as anything lesser than 0.2 resulted in more unmatched pairs. Values greater than 0.2 resulted in large imbalance between the treatment and control group. The results of the matching are shown in figure 11 . In order to estimate the causal effect, we performed a p test over the matched data. The output is shown in figure 12. As p value is close to 0, there is a difference between the means of both the groups, which proves our hypothesis that user's tweets are less likely to be depressive if there is engagement in the form of support over previous tweets.

8 CONCLUSION AND FUTURE SCOPE

Through the results of our experiments, we have shown how a collection of behavioral markers that appear on social media may be used to forecast postings that are suggestive of depression and, in turn, comprehend widespread depressive tendencies in populations. These estimates of depression can occasionally be used through our work and related research to assist early detection and prompt treatment of depression. Additionally, a major issue in public health

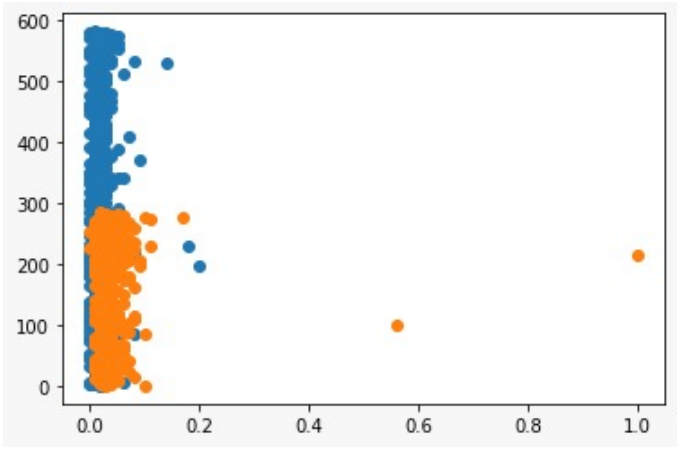

Figure 8: scatter plot of propensity scores between treatment and control users

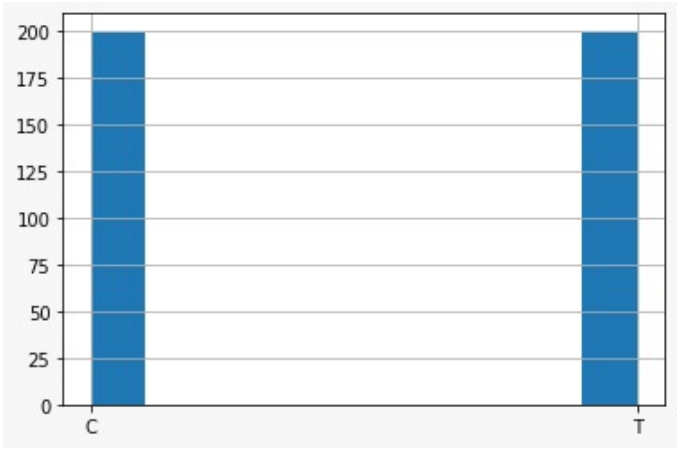

Figure 9: histogram of the matched users in the treatment and control groups

```
                        Score
Control/Treatment
C                   -1.216877
T                   -0.583646
The direct difference is -0.6332308831361724
```

Figure 10: Direct difference between average treatment effect before matching

is the under-reporting of a number of behavioral health issues. We anticipate that by using our suggested method, organizations will be better able to assess these issues and provide better healthcare.

When support is given, we have demonstrated causal links between the user and their social media (Twitter) activity. Additionally, we developed a useful method for estimating the number of support users receives when they post or tweet about their mental health and personal struggles. We gathered information from Twitter for

```
[1] "TABLE 1"
                        Stratified by controlortreatment
                         0                1                p      test SMD
  n                        584              286
  avg retweet (mean (SD)) 901.30 (2068.48) 550.88 (1347.39)  0.009        0.201
  avg reply (mean (SD))     0.28 (0.51)      0.70 (1.17)    <0.001        0.455
  avg likes (mean (SD))     2.08 (19.84)    10.28 (46.06)   <0.001        0.231
  avg quotes (mean (SD))    0.04 (0.40)      0.11 (0.67)     0.034        0.141
  avg length (mean (SD))   92.92 (50.20)   108.75 (46.80)   <0.001        0.326
  avg time (mean (SD))      2.57 (0.69)      2.58 (0.66)     0.865        0.012
[1] "Matched Data"
                        Stratified by controlortreatment
                         0                1                p      test SMD
  n                        287              287
  avg retweet (mean (SD)) 573.36 (1353.03) 548.97 (1345.43)  0.829        0.018
  avg reply (mean (SD))     0.58 (0.87)      0.69 (1.17)     0.185        0.111
  avg likes (mean (SD))     6.03 (38.89)    10.25 (45.98)    0.235        0.099
  avg quotes (mean (SD))    0.07 (0.57)      0.11 (0.67)     0.420        0.067
  avg length (mean (SD))  105.10 (46.35)   108.94 (46.82)    0.325        0.082
  avg time (mean (SD))      2.59 (0.65)      2.58 (0.66)     0.839        0.017
[1] "Matched Data with Caliper"
                        Stratified by controlortreatment
                         0                1              p      test SMD
  n                        112              112
  avg retweet (mean (SD)) 220.39 (531.74) 250.15 (524.28)  0.674        0.056
  avg reply (mean (SD))     0.33 (0.23)      0.34 (0.23)    0.790        0.036
  avg likes (mean (SD))     0.99 (1.00)      2.16 (1.79)   <0.001        0.808
  avg quotes (mean (SD))    0.01 (0.02)      0.01 (0.03)    0.355        0.124
  avg length (mean (SD))   89.67 (35.11)   89.94 (34.58)    0.952        0.008
  avg time (mean (SD))      2.58 (0.57)      2.58 (0.56)    0.993        0.001
```

Figure 11: Output after performing mahalanobis matching

```
        Paired t-test

data:  y_treatment_mdm and y_control_mdm
t = 1.9101, df = 111, p-value = 0.0587
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 -0.02945206  1.60313540
sample estimates:
mean difference
      0.7868417
```

Figure 12: Output of p-test to show the difference between means of both groups

over 872 users and then applied the BERT classification algorithm to identify a group of individuals who had been given a depression diagnosis.

We further divide the users into treatment and control groups, then we calculate the typical treatment impact for each group (ATE). We have also identified the study's intervention time. Propensity score and Mahalanobis matching were then used as two matching procedures, and the ATE was recalculated. This was done in order to improve the accuracy of the framework used to estimate the causal effect. These findings now support our causal inference hypothesis that "support" affects how many depressive-themed tweets a user sends after receiving the intervention.

Using this as our motivation we were able to achieve the following:

- Using Twitter scraping techniques, a dataset of 23,791 tweets from 870 different users was collected. This was accomplished with the Twinit was used to scrape user data from Twitter, and tweepy was used to scrape individual tweets.
- We used a statistical model (a transformer-based model classifier) to predict whether the message in the Twitter post is depressive or not. To classify the text, we used the deep BERT (Bidirectional Encoder Representations from Transformers) model.

- We proposed two new metrics, 'Support' and 'Score,' with 'Support' based on the tweet's engagement and 'Score' based on the difference between the percentage of depressive tweets from an individual user after the intervention and the percentage of depressive tweets from an individual user before the intervention.
- We used the variables to estimate the outcome effect and the Average Treatment Effect (ATE) and used various matching methods such as Propensity Score Matching, Mahalanobis Distance Matching, and Nearest Neighbor Matching.

After completing the aforementioned tasks, we believe we gained a better understanding of the causal relationships between users and their behavior on the social media platforms Twitter.

One of the limitations of our work is that we have done our analysis on a fairly smaller dataset, we believe we would be able to get better insight and understanding on the causal relationships between the behavior of the user on twitter and the support that they get when they tweet. Sentiment analysis is also a fairly young research topic so getting the perfect labels is always a problem as our classifier cannot detect sarcasm and humor. Another limitation could based on the training dataset we used for training our classifier, we assume that the dataset is accurate, which might not be the case as many times people in social media hence we are bound to missclassify some of those as depressive without context, so if there's any historically backed data where we are able to know that the person is depressive and their tweets are depressive and not sarcastic in nature, but this would possibly violate the privacy rights of the individual.

There are other methods that can be used to get causal inference, such as using structured causal learning, which can also provide us insights on the causal relationships between variables and the outcome.

More matching techniques can also be implemented to asses the result.

9 CODE

The datasets and the code for this project can be found at - https://github.com/nmehta32/Causality-between-support-and-depression

REFERENCES

[1] Abdel-Baki, A., Lal, S., Charron, D.-C., Stip, E., Kara, N. (2017). Understanding access and use of technology among youth with first-episode psychosis to inform the development of technology-enabled therapeutic interventions. Early Intervention in Psychiatry, 11(1), 72–76.

[2] Birnbaum, M. L., Rizvi, A. F., Correll, C. U., Kane, J. M., Confino, J. (2017b). Role of social media and the Internet in pathways to care for adolescents and young adults with psychotic disorders and non-psychotic mood disorders. Early Intervention in Psychiatry, 11(4), 290–295.

[3] Glen Coppersmith, Kim Ngo, Ryan Leary, and Anthony Wood. 2016. Exploratory Analysis of Social Media Prior to a Suicide Attempt. In Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology, pages 106–117, San Diego, CA, USA. Association for Computational Linguistics.

[4] Glen Coppersmith, Mark Dredze, Craig Harman, and Kristy Hollingshead. 2015. From ADHD to SAD: Analyzing the Language of Mental Health on Twitter through Self-Reported Diagnoses. In Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, pages 1–10, Denver, Colorado. Association for Computational Linguistics.

[5] Andrew Yates, Arman Cohan, and Nazli Goharian. 2017. Depression and Self-Harm Risk Assessment in Online Forums. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2968–2978, Copenhagen, Denmark. Association for Computational Linguistics.

[6] Elizabeth L. Ogburn, Oleg Sofrygin, Ivan Diaz, Mark J. van der Laan. Causal inference for social network data.

[7] Jared Jashinsky, Scott H Burton, Carl L Hanson, Josh West, Christophe Giraud-Carrier, Michael D Barnes, and Trenton Argyle. 2015. Tracking suicide risk factors through Twitter in the US. Crisis (2015).

[8] Hajime Sueki. 2015. The association of suicide-related Twitter use with suicidal behaviour: A cross-sectional study of young internet users in Japan. Journal of affective disorders 170 (2015), 155–160.

[9] Stice, Eric, Jennifer Ragan, and Patrick Randall. "Prospective relations between social support and depression: Differential direction of effects for parent and peer support?." Journal of abnormal psychology 113, no. 1 (2004): 155.

[10] Mclachlan, G.. (1999). Mahalanobis Distance. Resonance. 4. 20-26. 10.1007/BF02834632.

[11] Baltar, Valeria and Sousa, Clóvis Westphal, Marcia. (2014). Mahalanobis' distance and propensity score to construct a controlled matched group in a Brazilian study of health promotion and social determinants. Revista Brasileira de Epidemiologia. 17. 668-679. 10.1590/1809-4503201400030008.

[12] King, Gary, Richard A. Nielsen, Carter R. Coberley, James E. Pope and Aaron R Wells. "Comparative Effectiveness of Matching Methods for Causal Inference." (2011).

[13] Lateef Amusa, Delia North, Temesgen Zewotir,A tailored use of the mahalanobis distance matching for causal effects estimation: A simulation study,Scientific African,Volume 16,2022,e01155,ISSN 2468-2276, https://doi.org/10.1016/j.sciaf.2022.e01155.

[14] Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. Multivariate Behav Res. 2011 May;46(3):399-424. doi: 10.1080/00273171.2011.568786. Epub 2011 Jun 8. PMID: 21818162; PMCID: PMC3144483.

[15] Applying Propensity Score Methods in Clinical Research in Neurology Peter C. Austin, Amy Ying Xin Yu, Manav V. Vyas, Moira K. Kapral Neurology Nov 2021, 97 (18) 856-863; DOI: 10.1212/WNL.0000000000012777