Not Every Disability is Visible

ANALYSIS OF MENTAL HEALTH DISORDERS USING TWITTER

BURDE PRERANA KAMATH | JANANI MUPPALA | SRI MEGHA VUJJINI
ILS-Z639 SOCIAL MEDIA MINING | INDIANA UNIVERSITY BLOOMINGTON

Introduction

A mental illness is a condition that causes mild to severe disturbances in thought and/or behavior, resulting in an inability to cope with life's ordinary demands and routines. There are more than 200 classified forms of mental illness, some of the more common disorders being depression, bipolar disorder, dementia, schizophrenia, anxiety etc.

One in 5 adults experiences a mental health condition every year and the numbers are only rising. These illnesses aren't the result of one single event and research suggests multiple, linking causes like genetics, environment, and lifestyle.

In this project, we aim to narrow down the influence of **environment** on the mental health illness.

Based on the research *#WhyWeTweetMH: Understanding Why People Use Twitter to Discuss Mental Health Problems* [1], we use Twitter to extract the tweets talking about different mental disorders and analyze them.

Some of the previous work in this area includes but is not limited to: Predicting the onset of mental illness [2] where computational models were developed to predict the emergence of PTSD and Depression in Twitter users, analyzing the expressions for mental health on Twitter by checking for the magnitude of six mood dimensions [3] and different variations of sentiment analysis.

Most of the work done earlier talks about either the emergence or the existence of the mental health issues using Twitter data. We want to go a step further and predict, based on the location, where help is needed the most.

By analyzing the tweet text and drawing relationships per location, the aim is to make the process of assistance and recovery shorter and smoother.

Future Work

• The next step is to perform regression to find how much attention/help the illness is getting, at each location.

• Based on the final results, the project can be extended either ways:
   • Include more mental health diseases in the query and thus make sure the analysis is widened.
   • Find a reason for why maximum percentage of people have not mentioned their location while talking about their illnesses.
   • Find a relationship between lifestyle along with the environment and the mental health disorders.

Literature

1. #WhyWeTweetMH: Understanding Why People Use Twitter to Discuss Mental Health Problems, Natalie Berry et al., JMIR Vol 19(2017)

2. Quantifying Mental Health Signals in Twitter, Glen Coppersmith et al.

3. Analysis of Mental Health Expression on Twitter, Michael Billiot.

Methods

Data Collection

• Extracted data by querying the Twitter Search API
• Features: Author, Source, Location, Follower count, Retweet count, Friends count, Created at (time) and Tweet.
• Words queried: ADHD, Anxiety, Bipolar Disorder, BPD, Depression, Insomnia, OCD, PTSD, Schizophrenia, Suicide.

Data Preprocessing
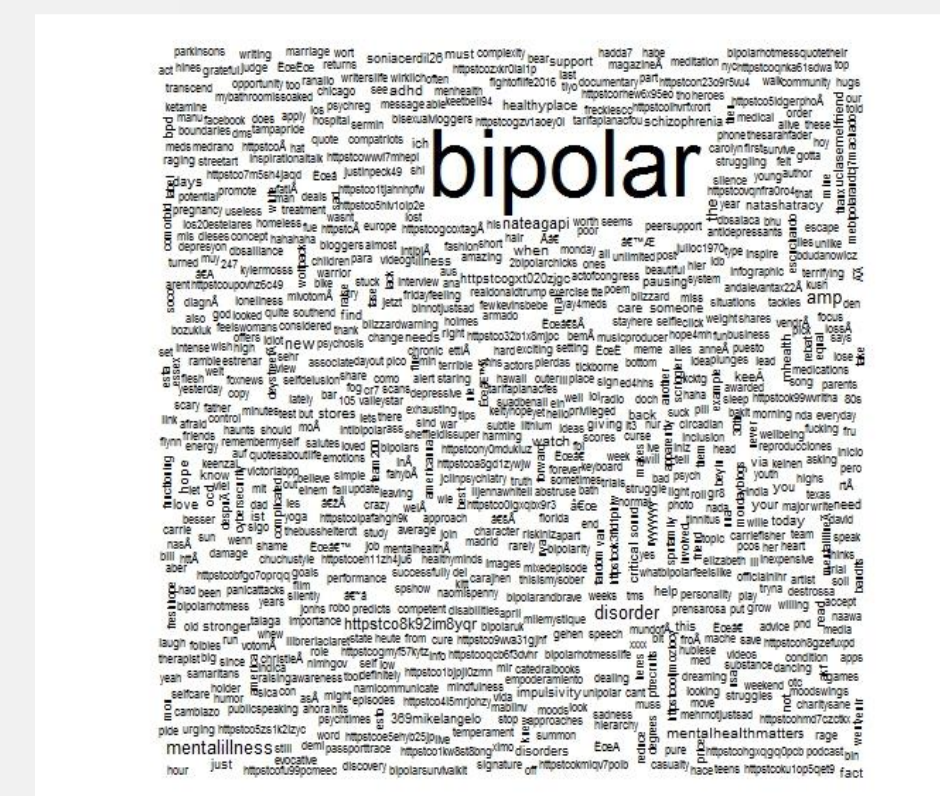
• The data extracted in a txt format encoded in UTF 8 for further preprocessing.
• Punctuations, cases, stop words and digits removed.
• Removed the retweets from the text.
• Using geographic locations, identify the areas with higher tweet rate.

Sentiment Analysis

• Performed this step to assess the severity of the issue.
• Used the AFFIN package in Python.
• Counted the negative words in the tweet text and assign a number.
• Inferred that higher the count, more severe the issue.
• For locations with higher count, perform further analysis for every disorder.

Linear Regression
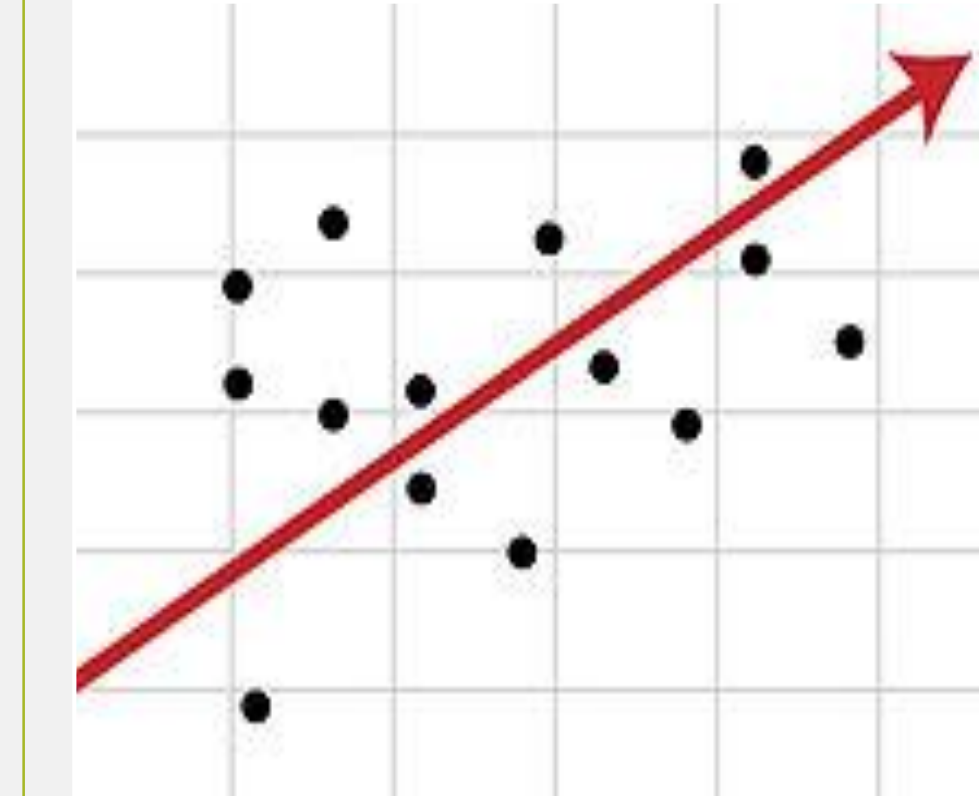
• Grouped the tweets based on its location feature.
• Perform Multiple Regression using the features retweet count and favorite count against location.
• This step is to predict the awareness of the issue and how much help is being given at every location.

Combining methods

• Combining sentiment analysis and regression analysis to get a picture of the issue per location.
• Sentiment analysis will give an idea of the severity of the issue while regression will give us an insight into the response at each location.
• Combining all of it to map it visually will make it easier for to provide assistance.

Results

The results obtained from the project so far are shown in the following two visualizations.
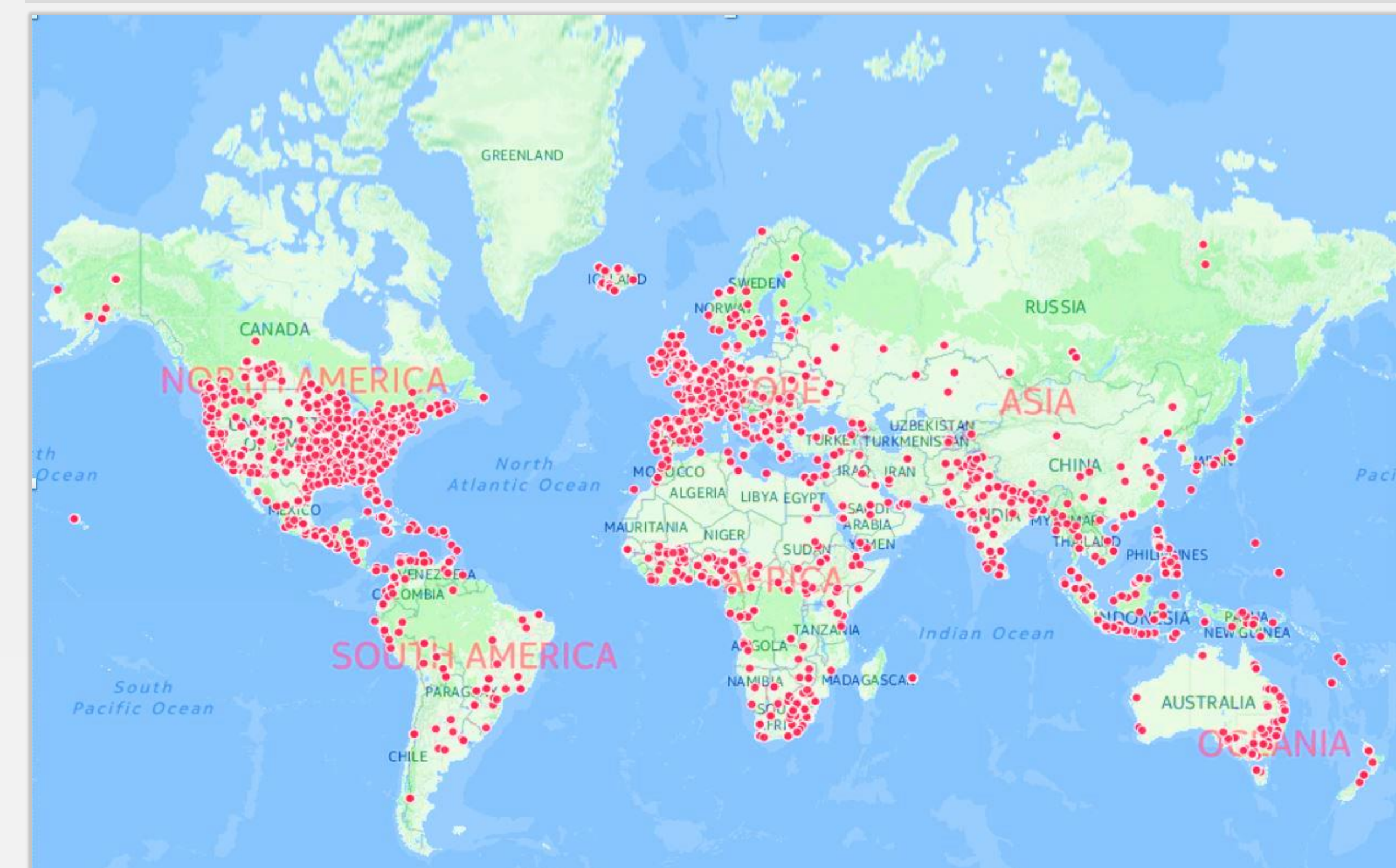
o Extracted all the tweets based on the query words.
o Cleaned the data as mentioned in the preprocessing step above.
o Mapped all the tweets based on their geographic locations and counted the number of tweets from each country.
o Countries with highest tweet counts chosen for further analysis: Australia, Canada, France, India, Malaysia, Philippines, United Kingdom, United States of America.

The bar graph is inverted because it's in the negative scale. We are only looking at sentiments with negative scores.

This chart shows the relationship between all the countries, their sentiment score for every mental illness disorder. In tabular format, it can be written as:

	Australia	Canada	France	India	Malaysia	Philippines	UK	USA
ADHD	-4	0	-5	-22	0	2	-25	-97
Anxiety	-41	-301	-122	-94	-92	-24	-224	-1271
Bipolar	-7	-7	45	9	-14	-1	25	-79
BPD	1	1	-12	3	-2	0	-40	-154
Depression	-25	-74	-51	-24	-2	2	-72	-513
Insomnia	-51	-79	-47	-144	-111	-28	-39	-448
OCD	6	-54	18	-6	-28	-2	-79	-84
PTSD	-61	-88	-56	-32	-11	-8	-41	-487
Schizophrenia	-4	-55	-21	3	-5	-3	-2	-153
Suicide	-86	-185	-66	-86	-54	1	-351	-690