From stage 2:

**Feedbacks:**

Legally: implication of obligation of the target customers

Consequences of false positives/negatives

Requirements of the dataset (training certificates etc…)

**Revised question: Can mental health disorder be predicted by social media activities?**

**Ethics**

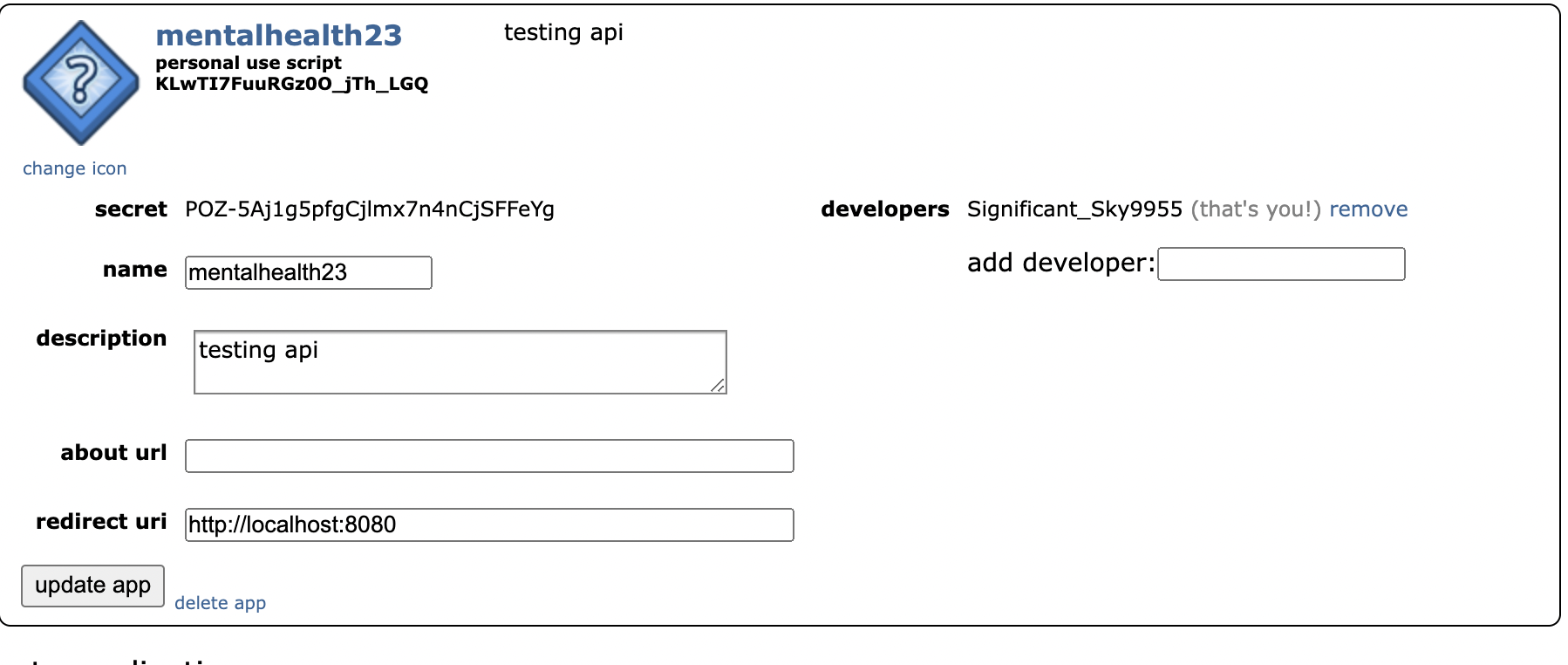
Ethics approval was not required for this study. The study follows research ethics guidelines for social media research; therefore, it is ethically compliant with the Institutional Review Board. The Market Research Society/Market and Social Research (Esomar) states that if it is public data there is no need for informed consent. The legitimation for processing public data is based on research for general interest, as identified in Papademas and IVSA, 2009: 254 and Papademas and IVSA, 2009: 255. The author(s) only compiled the identifying information needed for the study, and the data collected is not quoted directly in any part of this paper.

Instead using mental illness

Mental illness → mental health

Ethics requirement:

<https://authorservices.taylorandfrancis.com/editorial-policies/research-ethics-guidelines-for-stem-journals/>



Id: **KLwTI7FuuRGz0O\_jTh\_LGQ**

Secret:

| POZ-5Aj1g5pfgCjlmx7n4nCjSFFeYg  Significant\_Sky9955  <http://localhost:8080>  mentalhealth23 |
| --- |

**DataSet**

Dataset examples:

Reddit - mental health dataset (28 subreddit)

<https://zenodo.org/record/3941387>

<https://zenodo.org/record/4266616>

Twitter - self reported

<https://zenodo.org/record/5854911>

*Web Scraping for subreddit-* using hot or top ?

<https://vtiya.medium.com/how-to-scrape-reddit-data-in-2023-1613230f9e4d>

<https://towardsdatascience.com/scraping-reddit-data-1c0af3040768>

<https://medium.com/geekculture/a-complete-guide-to-web-scraping-reddit-with-python-16e292317a52>

<https://stackoverflow.com/questions/75384811/how-to-get-total-number-of-posts-of-a-subreddit-using-python>

<https://stackoverflow.com/questions/75678569/how-to-scrape-posts-from-a-subreddit-from-a-specific-year>

<https://stackoverflow.com/questions/75678569/how-to-scrape-posts-from-a-subreddit-from-a-specific-year>

Self reported diagnosis

<https://paperswithcode.com/dataset/smhd>

**Consequences of false positive/false negative:**

False positive:

False negative:

**Lecture notes**

Ecological fallacy is wrong

Generalizability - on average what your sample in your population looks like, not individual units

Sampling model- three steps:

identify pop- social media post about covid, particular kind of design

Fair sample - representative of pop,. Avg values → find results

Generalize back to pop - bring from theory and question

Problems with sampling model → can’t get the access, getting fair result is difficult

Proximal similar model - nearby, more likely to be our study or less likely to be our study

Gradient of similarity - ex. Next yr elections – but this yr elections - generalize we dont know if people change but ican generalize it generalize it with other person place or time certainty

Threats to external validity -

units: maybe skewed or biased - not just representative of pop,

Places: gathered from one place,

Times: reflect non-typical point time , ex. flu in the past and now

goodness of fit matters

Replication-

Role of replication - improve external validity- random assignment ex.

Proximity- understanding how your sample differs from your population of interest, compare the pop and sample mean, and make an argument how similar are they

Check Means and standard deviations same or different

Logic of underlying replication

Levels of replication-

Different types of replication

Ex. Classify two different types of things with same algorithm

Sth is similar enough? For replication → replication crisis: selective reporting-dont report the fail

Possibility of false positive or false negative

Data triangulation - limitation of replication as paradigm for external validity, multiple approaches to the question

Sampling: how do we draw the sample, what specific we get from sample

probability sampling- utilie some form of random selection,

Theoretical pop- pop whom you want to generalize, if you want to make a claim about whole pop, all social media users

Study pop- get access to, i could feasibly get data without limitation,

Ex. want to claim about social media users but access to fb and reddit

→ fb and reddit is my study and theoretical pop is social media users

Privacy restriction,

How you actually get access to? Different apis, services - fb- crowd tangle

Accessible pop- a subset of theoretical pop and measurable

Sampling frame- using twitter api, ml pipeline and find

Sample- subset of tweets collected between -

Data point- tweet

Sampling bias-

Sampling distribution- law of large numbers, central limit theorem- pop mean and sample mean on average will be normally distributed and how goood is the sample is,

Sampling error- gaussian distribution , how good is our sample, how sample match with population distribution

Simple random sample- each unit equally randomly selected from population

numpy.rand()

stratified random- split into sub pop, sample uniformly from each pop,

Class balancing in machine learning

High quality, med quality, low quality, unrated for example

Stratified : Good for how different from one another

Simple random: good for population as a whole

Cluster random sampling: twitter dataset- json file by dates, 1000 tweets from each day

Multi stage sampling: given large num of tweets, look at N tweets total, randomly select set of d days, within each day classify all tweets about flu or not flu and select equal num of each category

Other sampling: Convenience sample- get alot of data, purposive sampling- certain keywords or urls,

**Questions:**

1. **Identify a source of pilot data that you will use to answer your question**

* Reddit dataset
  + Web scraping - 2023 top posts - mental health related threads:
    - Mental health, depression, anxiety, bipolar, suicideWatch, mental illness, adhd

1. **Present descriptive statistics of these data, including a list of samples, features, measures of central tendency - mean, median or mode as appropriate**

* samples:
* Features:
* Measures of central tendency: mean -

1. **What is your unit of analysis, and why - We discussed this last time, but now rethink it based on your sampling plan**

* Unit of analysis: subreddit posts/comments related to mental illness - mental health, mental illness, adhd, depression, bipolar, suicide, and anxiety.

1. **What is the theoretical population from which this unit will be selected? Why?**

* All reddit users post/comment on mental health-related subreddits at any given time
  + We want to generalize reddit user’s behavior within mental health related subreddits

1. **What is the accessible population from which this unit will be selected? Why?**

* Top post/comment on subreddits related to mental health on 2023
  + We don't have a dataset of other years
* How you get access: Reddit is a public website -> using web scraping to collect data directly from the site

1. **What is the sampling frame that will you use to select your sample? Why?**

* Sampling frame: using pipeline of classification of Natural Language Processing with keywords people using in mental health subreddits
* Sample: subsets of mental health related reddit post/comment dataset on 2023
* Level of replication: different subreddit groups with same NLP algorithms
* Data point: words from posts/comments

1. **How large will your sample be? Why?**

* Sample is subset of mental health related reddit post/comment dataset in 2023
* Sampling bias: only some specific mental health-related subreddits are scraped -> not a representative of all reddit users or everyone with mental health disorders

1. **What are some threats to the external validity of your sampling strategy?**

* Units: Our sample might not be representative of the population since the data is restricted to very specific subreddits
* Place: restricted to reddit - gathered from one place
  + Reddit has specific characteristics that allow people to be more straightforward and comfortable expressing their opinions/feelings. People truly get the pros of being anonymous on here
* Times: reflect to 2023 - might affected by covid

1. **Will you use probability sampling or nonprobability sampling? Why?**

* Probability - our goal is to generate a representative sample from a population (sample: subreddits related to mental health, population : )

1. **10.Within your choice to #9, what specific sampling strategy will you use? Justify this by relating it back to your research question.**

* Random sampling, straified???
* We chose a random sample from