From stage 2:

**Feedbacks:**

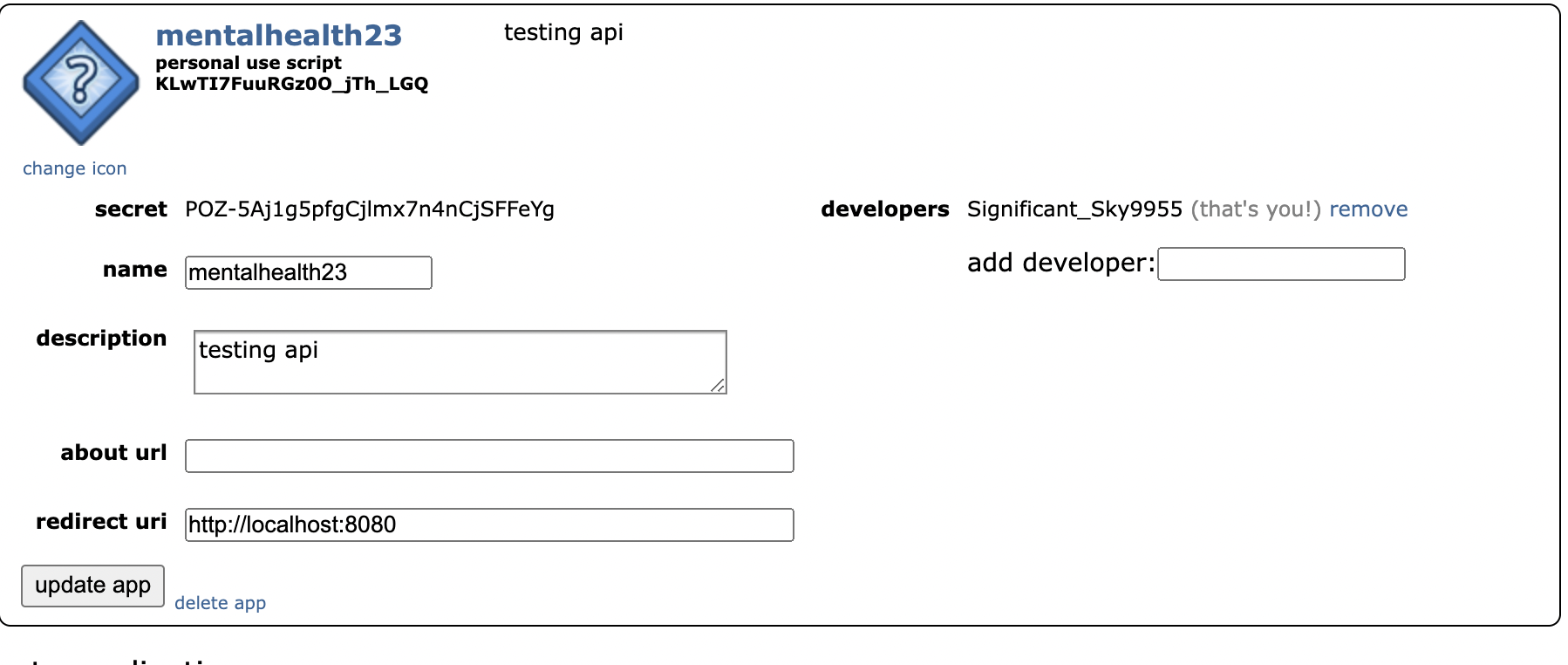
Legally: implication of obligation of the target customers

Consequences of false positives/negatives

Requirements of the dataset (training certificates etc…)

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>

→ statistics : using score, upvote ratio, total comments, frequency of ID posts

→ NLP: master data - title and post text separately (different num)

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

False positive:

False negative:

**Lecture Notes**

* Construct validity

What are you measuring,

Cause and effect → data rlly good measure,

**•Predictive Validity:** Does it predict what it should be able to predict (Twitter predict the flu)? •Concurrent Validity: Does it help us to differentiate between two things that are different (influenza awareness vs. influenza infection)?

•Convergent Validity: Is it similar to other things that measure the same thing? (ILI vs. Twitter mentions)

•Discriminant Validity: Is it different from other things that measure different things? (Twitter mentions vs. Zika)

* Definition of constructs and construct validity

Construct - theoretical abstraction to operationalized - can measure

Example: quality of restaurant → average num of stars in online review

Definitionalist - precisely essential measure

Relationalist - define similar concept

* Convergent and discriminant validity

“Measures of constructs that theoretically should not be related to each other are, in fact, observed to not be related to each other (that is, you should be able to discriminate between dissimilar constructs)”

Things that are similar is converge each other

Scale measure -

Discriminant validity - measures of constructs that theoretically should not be related to each other

* Measuring comprehensibility

Make correlation table, PCA table - measure of text comprehensibility (kaiser’s criterion)

Pca is measuring the readability - composit metric all of the measures and drives

* What is the convergent - similar concept from the validity
* What is the discriminant - differences from the validity
* Threats to construct validity

Not clear on what you are measuring

Mono-operation bias - only use one measure

Interaction of different treatments - selection bias, sample is exposed to other factor

Interaction of treatment and testing - response bias , the way that you measure affects the result

Restricted generalizability - FORGOT THE MEASURE important

Confounding constructs with level of constructs - measuring range is too narrow

Social threats of construct validity

* hypothesis guessing
* Evaluation apprehension
* Experiment expectancies
* Multi-trait multi method matrix

Reliability: monotrait monomethod

Validity: monotrait heteromethod

What about the others?

Trochim argues they aren’t necessary for construct validity

* Reliability

True score theory

* Why measures have low reliability?
* According to true score theory the noise overwhellms the signal
* Observed sore = true abilit + random error

Measurement error

* Relation to bias-variance tradeoff
* Levels of measurement and scaling
* Unobtrusive measures and ethics

**Feedback from stage 3**

So probability sampling?

Limitation - the limitations. What I don't understand is why you're only looking at those

What does "top posts" mean?

* Top rated posts from 2021 to 2023

You're going back and forth between datapoints being words, unit of analysis are users, and sample being posts

* Datapoint : words -
* Unit of analysis:
* Sample:

relation between posts and users is not at all clear

You didn't really justify why you're looking at the post when your population is users

It seems like you're collecting a lot of metrics, but are you indexing the content at all?

The features that you present are hard to interpret; good that you see the histograms because that definitely suggests a statistical analysis approach

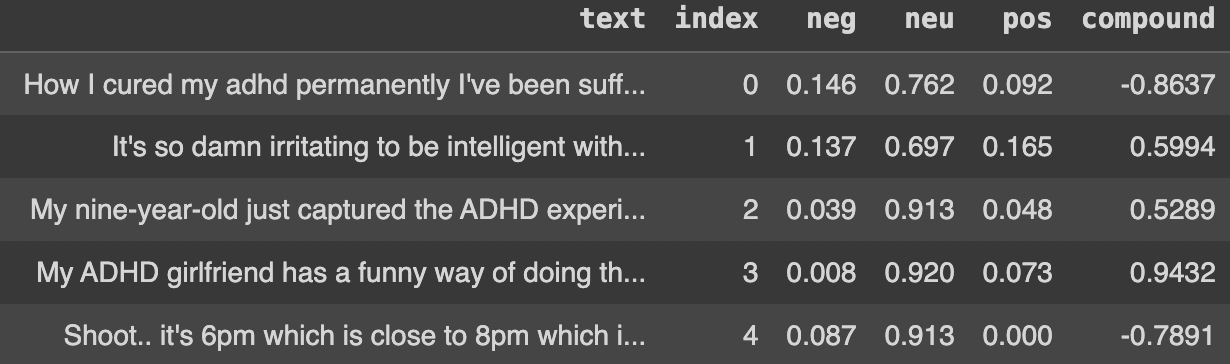
Construct 1 - mental wellbeing

Construct 2- text data

Construct 3 - quantitative data

Hypothesis - mental well being

**Title & Post text sentiment analysis**



Using nltk sentiment vader library to analysis the text contains polarity score of each negative neutral or positive by words in the text (vader use each words if they are pos, neg, or neu)

we decide to use threshold of 0.2 for sentiment analysis

Later on, we are moving forward to compare different subreddit group and their score and sentiment analysis getting a result of accuracy of positive and negative posts

Ex.

**Positive accuracy = 79.93450431676095% via 3359 samples**

**Negative accuracy = 57.75229357798165% via 2180 samples**

**Questions**

1. **Revisit research questions. Restate the question as a hypothesis, and list the key constructs and any proposed relationship (causal or otherwise) between them.**

Revisit research questions: Can users’ mental health illness be predicted by negative expressions/behaviors on social media?

Restate the question as a **hypothesis**:

1. **By examining users' behavior on social media, we can detect their potential susceptibility to experiencing mental health disorders**
2. Violent or negatively expressive behaviors on social media can indicate underlying real life mental illnesses

Key constructs and any proposed relationship between them:

Constructs:

* Construct 1: Frequency of reddit usage (User ID)
* Construct 2: Readability of posts (text classification, Automated Readability Index)
* Construct 3: Emotionality of posts (negative/positive accuracy of text)
* Construct 4: Accuracy of predictability of mental illness

Predictive validity

Causal Relationships:

* Construct 1, 2 and 3 cause 4
  + Frequency of usage, text classification (readability index), and sentimental analysis of text indicates to predict mental illness of reddit users

1. **Revisit pilot data. List all of the features in your pilot data. Create a table relating your features to your constructs. Explicitly highlight those constructs for which you don’t have features, and those features that don’t correspond to constructs in your question**

List all of the features in pilot data:

* Title -> 1
* Post text -> 1
* User ID -> 1
* Score -> 2
* Upvote Ratio -> 2
* Total Comments -> 2
* Created On
* Post URL
* (More features to be added after NLP)

Table related features to your constructs:

| Construct 1: frequency of reddit usage | Score, User ID |
| --- | --- |
| Construct 2: Readability of posts | Title, Post text, total comments, upvoted ratio |
| Construct 3: Emotionality of posts | Title, Post text |
| Construct 4: prediction on mental illness | Construct 1 to 3 |

No need: post url, created on

New feature after sentiment analysis: compound with threshold of 0.2 → accuracy of pos or neg on each post

1. **For each construct missing a feature, come up with a plan to measure that construct (or otherwise develop a scale to measure that construct)**

Construct 4: no data, will be predicting it using the pre-existing features that are used to measure construct 1, 2 and 3 and add new feature of sentiment analysis

1. **For each feature, indicate whether it is nominal, ordinal, interval, or ratio. Use your constructs to justify these selections.**

* Title: nominal
* Post text: nominal
* User ID: nominal
* Score: ratio
* Upvote Ratio: ratio
* Total Comments: interval

Nominal- back number of football

Ordinal- ordered, this data < that data,ordered logistic regression, class each items

Interval - distance, continuous

Ratio - absolute zero , temperature (klevin is only)

1. **Use your pilot data to demonstrate as much validity of each feature/construct pair as you can. For each feature/construct pair, justify why this is a valid mapping. Specifically indicate whether it has face validity, content validity, predictive validity, and/or convergent-discriminant validity**

Predictive Validity: Does it predict what it should be able to predict (Twitter predict the flu)? •Concurrent Validity: Does it help us to differentiate between two things that are different (influenza awareness vs. influenza infection)?

•Convergent Validity: Is it similar to other things that measure the same thing? (ILI vs. Twitter mentions)

•Discriminant Validity: Is it different from other things that measure different things? (Twitter mentions vs. Zika)

Content validity : keywords commonly used

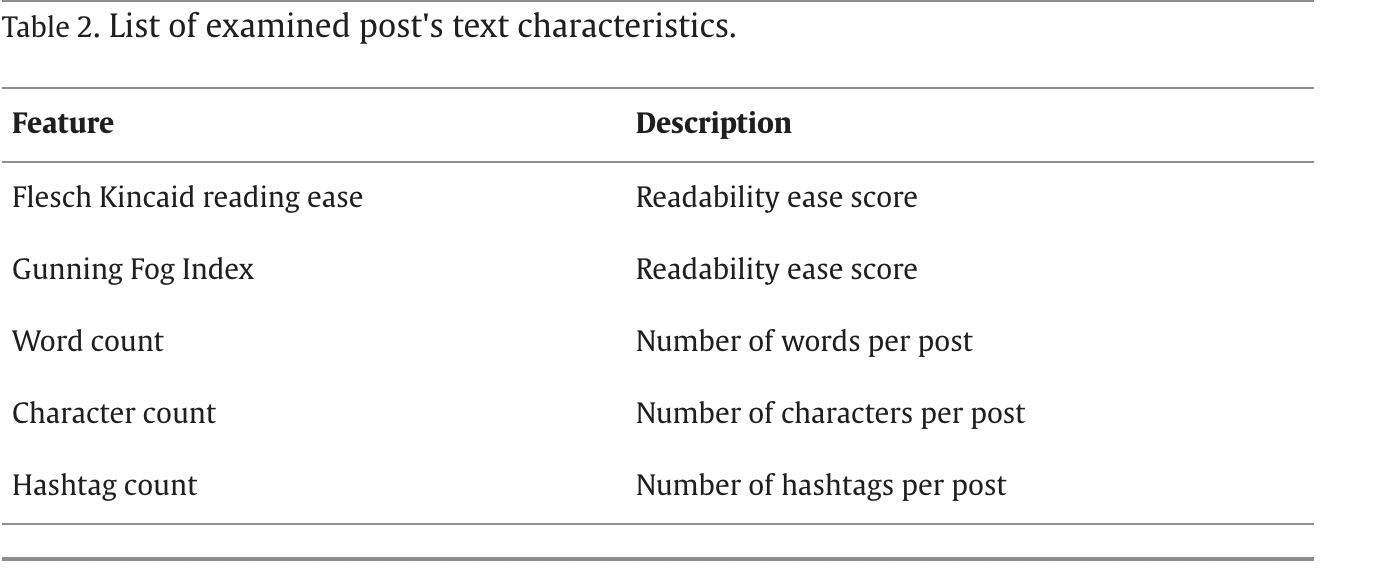
concurrent validity : associated with outcome

Convergent: sentiment lexicon

Predictive: using random forest, logistic model,

Every construct are having face validity

| Construct 1: frequency of reddit usage | Score, User ID | Content, face   * Frequency of usage shows how much users contribute to the reddit also means if their posts are appropriate to use as “reddit users with mental illness” instead of posting just once |
| --- | --- | --- |
| Construct 2: Readability of posts | Title, Post text, total comments, upvoted ratio | Concurrent validity, content validity, face  Text length, language and vocab, sentence structure, punctuation and grammar, context and tone,accessibility, engagement   * Title and post text for measuring factors of text length, language and vocab, sentence structure, context and tone, basically these will be measurement of how text characteristics impact user engagement * Total comments, upvoted ratio are engagement measurement |
| Construct 3: Emotionality of posts | Title, Post text | Convergent-discriminant validity, content validity, face   * Measuring either pos or neg accuracy of posts if sample posts are negative to measure mental illness- we expect negative for mental illness * sentiment analysis using nltk sentiment analysis |
| Construct 4: prediction on mental illness | Construct 1 to 3 | Predictive validity, face  Valid bc it predicts what it should be - predicting mental illness by sentiment analysis and readability of text by users who are likely to frequently visiting mental health related subreddit |



* Text Length: The length of a social media post can impact readability. Short and concise posts are typically easier to read and understand. Longer posts may require more cognitive effort to comprehend, and readers might lose interest quickly.
* Language and Vocabulary: The choice of words and vocabulary in social media posts can influence readability. Posts with simple and commonly used words are often more readable, while posts with complex or jargon-heavy language may be less accessible to a broader audience.
* Sentence Structure: The structure of sentences can affect readability. Short and straightforward sentences are easier to follow, whereas long, convoluted sentences can make text more challenging to understand.
* Punctuation and Grammar: Proper punctuation and grammar are essential for readability. Posts with frequent errors in punctuation and grammar can be difficult to read. For example, excessive use of exclamation marks or emojis can impact the perceived readability.
* Visual Elements: Social media posts often include visual elements like images, videos, and formatting (e.g., bullet points or headings). These visual aids can enhance the readability of posts by breaking up text and making it more engaging.
* Context and Tone: The context and tone of a social media post can affect readability. Posts with a clear purpose, message, and a tone appropriate for the intended audience are more likely to be readable.
* Accessibility: Consider the accessibility of social media posts for individuals with disabilities. Ensure that posts are accessible to people with visual or hearing impairments by providing alternative text for images and captions for videos.
* Engagement: High levels of user engagement, such as likes, comments, and shares, can indicate that the content is engaging and readable. Analyzing these metrics can provide insights into how readers perceive the post's readability.
* To quantitatively assess the readability of social media posts, you can use readability formulas and metrics like the Flesch-Kincaid Readability Index, Gunning Fog Index, or Coleman-Liau Index. These tools analyze factors such as sentence length and syllable count to assign a readability score to the text.

1. **For each feature/construct pair, test its convergent-discriminant validity using the data that you have available. If you can’t do so using your available data, develop a plan to gather data to do so, and indicate the resources that would be required to implement this plan. If you can, implement this plan. If you can’t, justify your continued use of this feature.**

| Construct 1: frequency of reddit usage | Score, User ID |  |
| --- | --- | --- |
| Construct 2: Readability of posts | Title, Post text, total comments, upvoted ratio |  |
| Construct 3: Emotionality of posts | Title, Post text |  |
| Construct 4: prediction on mental illness | Construct 1 to 3 |  |

For emotionality, we will perform sentiment analysis with NLP techniques to gather data for this construct. The expected features are: Negative scores, negative scores, neutral scores and compound sentiment scores (a metric that combines the positive, negative and neutral scores to provide an overall sentiment intensity) → **convergent when posts are negatively correlated and discriminant when there are any correlations or more positive posts**

For readability of context, we will quantify the text from the data by using metrics like Flesch-Kincaid Readability Index, gunning fog index, or coleman-liau index, and automated readability index, sentence length, lexicon (num of sentence i think) to assign a readability score; and using existing feature of total comments and upvoted ratio to measure how people engagement with posts

→ **convergent when readability is when there are PCA and**

For frequency of usage, we will use existing score and user id to see how frequently users engage in reddit by actively post or comment - num post on mental illness related group and score as a reddit user

→ **convergent when our user id have high score,discriminant when user id does not have frequently post or having high score**

* readability formulas and metrics like the Flesch-Kincaid Readability Index, Gunning Fog Index, or Coleman-Liau Index. These tools analyze factors such as sentence length and syllable count to assign a readability score to the text.

1. **For each feature, justify why this is a reliable measure by quantifying measurement error. If you cannot do so,come up with a plan to collect data to do so.**

**Mono operation- data reduction**

**Social threat construct validity - using human data**

**Inter rater - multiple different measurement or test - sat or gre or multiple people amtrak enoote data**

**Test retest0 same test over over second hand smoke everyday**

**Parallel - two different samples and measure same thing**

**Internal consistency - doing same test giving someone sat writing verbal math component .. topic model to tweet posts different subreddit category and see if the result is same**

**Continuous is correlation categorical is percent**

* Title:
  + Justification: Title of a Reddit post is typically a direct reflection of the content or topic being discussed. However, there can be some degree of ambiguity or subjectivity in interpreting titles.
  + Measurement Error Quantification:
  + Post Text:
    - Justification: The post text provides detailed information about the content of the post. However, measurement error can arise from variations in writing style, grammar, and the presence of slang or abbreviations.
    - Measurement Error Quantification: Similar to titles, an inter-rater reliability study can be conducted to assess agreement between raters when categorizing or analyzing post text.
    - User ID:
    - Justification: User IDs are unique identifiers for individual Reddit users. They are typically reliable as long as they are collected accurately and consistently.
    - Measurement Error Quantification: The measurement error for user IDs is likely to be very low, assuming that the data collection process is accurate and there are no issues with duplicate or incorrect IDs.
    - Score:
    - Justification: The score of a post represents the net upvotes it has received, providing an indication of its popularity and relevance within the community. However, there can be variability due to factors like vote manipulation or algorithm changes.
    - Measurement Error Quantification: The measurement error for scores can be estimated by comparing scores of posts over time and assessing the degree of variability. Additionally, conducting a reliability analysis on score data across multiple observers can help quantify any discrepancies in recording scores.
    - Upvote Ratio:
    - Justification: The upvote ratio indicates the proportion of upvotes to total votes, giving an idea of the community's sentiment towards a post. It is generally reliable, but may be affected by vote brigading or manipulation.
    - Measurement Error Quantification: Similar to score, the measurement error for upvote ratio can be estimated by analyzing its variability over time and conducting inter-rater reliability studies if applicable.
    - Total Comments:
    - Justification: The total number of comments on a post reflects the level of engagement and discussion it has generated. However, measurement error may occur due to factors like deleted comments or delayed updates in the comment count.
    - Measurement Error Quantification: Analyzing the consistency of comment counts over time and comparing them with archival data can help estimate measurement error. Additionally, conducting an inter-rater reliability study for comment counting can be valuable.

1. **For each feature, indicate any potential sources of methodological bias. Come up with (and if possible, implement) a plan to mitigate this bias.**

* Title: Titles may not accurately represent the content or may contain triggering/sensational language that attracts specific types of posts.
* Post text: Length and content of post text may vary widely

> mitigate: Conduct a content analysis of titles to identify and control for any biased language

* User ID:

> mitigate

* Score: Popular or controversial posts may receive a higher score

> mitigate: Normalize scores by the total number of votes to account for variations in post popularity; Use statistical techniques (e.g., regression analysis) to control for the influence of score on the analysis

* Upvote Ratio: Different communities on Reddit may have varying norms and expectations for upvoting

> mitigate: Include subreddit as a categorical variable in the analysis to account for community-specific biases (did this); Consider using the upvote ratio relative to the subreddit it belongs to, rather than an absolute measure.

* Total Comments: Controversial topics may generate more comments

> mitigate: Normalize total comments by post score or views to account for variations in post popularity; Consider including a sentiment analysis of comments to capture user engagement with the post (trying to do this)

1. **For each feature/construct pair, indicate the appropriate measures of reliability: Cronbach’s alpha, internal consistency, Fleiss’ Kappa, etc. All of these measures depend on multiple features per construct.If you don’t have this for some constructs, develop a plan to gather data to do get the appropriate features, and indicate the resources that would be required to implement this plan. If you can implement this plan, do so.**

**Cronbach’s alpha**

Interval or ratio data

Split half reliability

Divide items (in our case features ) into two and see which item to drop to improve reliability

Using r, it will tell you when item is drop what is the correlation

**Similar to Internal consistency**

Basically find the correlation of all of the items that are measuring the same construct

Ex. psychological test measuring depression or intelligence - ordinal and nominal

**→ internal consistency (correlation of all item from same construct) → cronbach (divide items into two which are to measure same construct)**

**Internal consistency and cronbach’s ~ so basically two or more independent variable and one dependent**

**Cohen’s k**

Nominal data, measure between two dependent categorical samples

Two observers independently observed and record behaviors at same time

Range 1 to -1 and negative kappa means observed level of agreement is less than what you expect by change

**Ex. two bankers to classify 100 customers in two classes for credit rating - good and bad based on creditworthiness**

**Fleiss ‘ Kappa**

Expansion on cohen’s which more than two dependent categorical samples

Provide categorical judgement like content analysis or medical diagnosis

**Cohen and fleiss ~ two or more independent and two or more dependent**

1. Inter-Rater Reliability: Consistency across raters (observer)

2. Test-Retest Reliability: Consistency across time

3. Parallel-Forms Reliability: Consistency between two measures derived in the same way from different samples of the same data

4. Internal Consistency Reliability: Consistency between items in the same test

| Construct 1: frequency of reddit usage | Score, User ID | Internal consistency   * Correlation of all time |
| --- | --- | --- |
| Construct 2: Readability of posts | Title, Post text, total comments, upvoted ratio | Cronbach   * Divide items (supposedly title&post versus total comments&upvoted) and see to measure same construct |
| Construct 3: Emotionality of posts | Title, Post text | Cohen’s   * Text as independent * Neg or pos sentiment result as dependent |
| Construct 4: prediction on mental illness | Construct 1 to 3 | Fleiss kappa   * We have several independent and several dependent * Provide categorical judgement (our is content analysis i think) * Relevant for using different construct (different categories) to predict construct 4 |

1. **For each existing \*and proposed\* feature and construct, list the threats to construct validity for your study and indicate any potential source of ethical concerns that may arise by collecting your data. What is your plan to fix these**

**Construct 1: Frequency of reddit usage:**

* Privacy Concerns: Collecting data on participants' online activities, even within mental health-related subreddits, may intrude on their privacy.

> Plan: Include a clear disclaimer about privacy and data usage in the informed consent process. Ensure data is anonymized and stored securely

**Construct 2: Readability of posts:**

* Subjective Interpretation: Assessing readability is subjective; different raters may have different opinions.
* Content depends on context, not just implication
* Ethical concerns: may reveal users' sensitive information, especially in posts related to mental health.

> Plan: Utilize established readability metrics like Flesch-Kincaid or Gunning Fog, and consider multiple raters for assessment.

Ensure that posts are analyzed in their original context and not taken out of context.

**Construct 3: Emotionality of posts**

* emotional expressions can be subtle and context-dependent > subjective
* Some posts might contain mixed emotions, making it harder to assign a single emotional label > ambiguity

> Plan: Use established sentiment analysis tools and techniques to quantify emotions in posts.

Consider using more advanced methods (e.g., machine learning classifiers) to capture nuanced emotional expressions.

**Construct 4: Accuracy of predictability of mental illness**

* Limited Data: Predicting mental illness based on online posts might not capture the full complexity of mental health conditions.
* Ethical Concerns: Predicting mental illness could lead to unintended consequences, such as stigma or inappropriate interventions.

> Plan: Use a multi-modal approach, considering various data sources (e.g., self-reports, linguistic analysis, behavioral patterns).