Proceedings of the Eighteenth International AAAI Conference on Web and Social Media (ICWSM 2024)

Reliability Analysis of Psychological Concept Extraction and Classification in User-penned Text

Muskan Garg[1], **MSVPJ Sathvik**[2], **Shaina Raza**[3], **Amrit Chadha**[4], **Sunghwan Sohn**[1]

[1] Mayo Clinic, Rochester, MN, USA,
[2] IIIT Dharwad, Karnataka, India,
[3] Vector Institute, Ontario, Canada,
[4] IIT Kharagpur, India,
garg.muskan@mayo.edu, 20bec024@iiitdwd.ac.in, shainaraza@vectorinstitute.ai,
amritchadha66@gmail.com, sohn.sunghwan@mayo.edu

Abstract

The social NLP research community witness a recent surge in the computational advancements of mental health analysis to build responsible AI models for a complex interplay between language use and self-perception. Such responsible AI models aid in quantifying the psychological concepts from user-penned texts on social media. On thinking beyond the low-level (*classification*) task, we advance the existing binary classification dataset, towards a higher-level task of reliability analysis through the lens of explanations, posing it as one of the safety measures. We annotate the *LoST* dataset to capture nuanced textual cues that suggest the presence of low self-esteem in the posts of Reddit users. We further state that the NLP models developed for determining the presence of low self-esteem, focus more on three types of textual cues: (i) *Trigger*: words that triggers mental disturbance, (ii) *LoST indicators*: text indicators emphasizing low self-esteem, and (iii) *Consequences*: words describing the consequences of mental disturbance. We implement existing classifiers to examine the attention mechanism in pre-trained language models (PLMs) for a domain-specific psychology-grounded task. Our findings suggest the need of shifting the focus of PLMs from *Trigger* and *Consequences* to a more comprehensive explanation, emphasizing *LoST indicators* while determining low self-esteem in Reddit posts.

Background

Mental disorders are a significant contributor to global mortality rates, accounting for approximately 14.3% of deaths worldwide (Walker, McGee, and Druss 2015). According to *the Global Burden of Diseases, Injuries, and Risk Factors Study 2019*, mental disorders continue to rank among the top ten causes of burden globally, without any indication of a reduction in their impact since 1990 (Collaborators et al. 2022). In the past few years, extensive investigations have illuminated the intricate associations between low self-esteem and various mental disorders, as evidenced by the Diagnostic & Statistical Manual of Mental Disorders (DSM-5) (Rouault et al. 2022; Cella et al. 2022).

Psychology-Ground. The notion of *low self-esteem* has been identified as a predisposing factor for the onset of social anxiety (Acarturk et al. 2009). In line with this, scholarly evidence underscores the pivotal role of low self-esteem in heightening the susceptibility to depression, anxiety, suicidal tendencies, impaired cognitive performance, compromised sleep quality, and diminished overall health (Korkmaz, Korkmaz, and Çakar 2019). Additionally, young individuals with recurrent suicide attempts exhibit a pronounced tendency towards enduring suicidal thoughts, challenges in interpersonal relationships, feelings of detachment, and diminished self-efficacy (Choi et al. 2013). Historical studies have shed light on the diminished self-assurance prevalent among individuals with low self-esteem, and its consequential association with reduced social involvement (Watson and Nesdale 2012). Emphasizing the paramountcy of *high self-esteem* as a foundational human requisite, a distinguished American psychologist has delineated two distinct variants of "esteem" within his seminal theory on the *'hierarchy of human needs'*: (i) the *external validation from peers*, encompassing acknowledgment, accomplishments, and respect, and (ii) the *intrinsic self-regard*, encapsulating self-affection, assurance, capability, and prowess (Cox 1987).

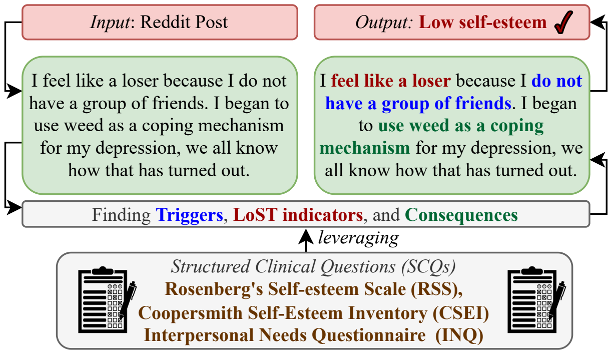

Input: Reddit Post
Output: Low self-esteem ✓
I feel like a loser because I do not have a group of friends. I began to use weed as a coping mechanism for my depression, we all know how that has turned out.
I **feel like a loser** because I **do not have a group of friends**. I began to **use weed as a coping mechanism** for my depression, we all know how that has turned out.
Finding **Triggers**, **LoST indicators**, and **Consequences**
leveraging
Structured Clinical Questions (SCQs)
Rosenberg's Self-esteem Scale (RSS),
Coopersmith Self-Esteem Inventory (CSEI)
Interpersonal Needs Questionnaire (INQ)

Figure 1: The overview of our task. We annotate the textual cues indicating the low self-esteem aspect in human-writings, emphasizing the need of focusing LoST indicators more than triggering words and final consequences.

The correlation between low self-esteem and 21 distinct disorders is well-documented, encompassing diagnostic criteria, associated features, risk factors, and consequences specifically linked to individuals grappling with low self-esteem (Kolubinski et al. 2018). Individuals often share their feelings and thoughts on social media, especially when reflecting on personal experiences related to low self-confidence. Identifying and understanding the deep connection between low self-esteem and related disorders is crucial for developing effective interventions and promoting mental health.

Reliability Analysis: Motivation. In the nuanced arena of analyzing user-penned content on social media for indicators of low self-esteem, reliability is paramount. As mental health practitioners and platforms increasingly recognize the value of insights from social media to tailor interventions, the role of trustworthy models becomes even more critical. Thus, the intersection of mental well-being and technological analysis demands both accuracy and trustworthiness. Reliable analysis ensures that the textual cues being flagged as indicative of low self-esteem are genuine, eliminating false positives and ensuring that genuine cases don't go unnoticed. By reliably interpreting an individual's online expression, there's an opportunity to provide more personalized guidance and resources. For such analytical models to be wholeheartedly embraced by both the professional community and the users themselves, they must be deemed reliable.

Our Contributions. The current landscape offers abundant open-source datasets for research and application development across various fields. Yet, there exists a notable dearth of high-caliber psychology-grounded datasets, especially pronounced in the ever-evolving realm of healthcare informatics and NLP.Datasets sourced from participatory healthcare informatics frequently exhibit attributes of intricacy, randomness, and an absence of clear structure. The performance of models is invariably tied with the integrity, and applicability of the dataset. With these considerations, a paramount challenge in psychology-driven dataset compilation is to ensure the consistency and reliability of the annotations. To navigate and simulate this intricate landscape, we choose to design an annotation scheme to discern textual-cues for psychological concept (low self-esteem) extraction and classification in user-penned text. The assembly process of our annotation scheme and annotated dataset reveals discernible gaps in content comprehensiveness and mark its adaptability for healthcare utilization.

Building upon this foundational knowledge, annotation guidelines were meticulously curated through the joint endeavors of two specialists: (i) a senior clinical psychologist and (ii) a social NLP expert. Our work enables the development of reliable NLP models in the near future, to facilitate *higher-level tasks* (Sheth et al. 2021)(Raza and Schwartz 2023). As such, we contribute by advancing our studies with the responsible classification of user-penned text through the lens of three types of textual cues: *Trigger, LoST indicators, Consequences*, collectively we mention them as TLC. To this end, we define this task as the first of its kind to detect appropriate textual cues that aid in decision-making of low self-esteem detection. We define these textual cues as set of words indicating ”explainability” and are focused by the attention mechanism of classifiers while making decisions.

Corpus Construction

Corpus Collection We construct a dataset for identifying text-spans reflecting low self-esteem, from the collected instances from subreddits `r/depression` and `r/suicidewatch` from 2 December 2021 to 4 January 2022. By using this curated dataset, we aim to ensure the quality and relevance of the data for our specific research objectives. We further perform the manual cleaning and removed all nearly empty posts (posts having less than 10 words) and posts containing only URLs to accommodate the subjectivity of the task. We included the user-penned experiences and excluded all the general posts borne out of concerns for fellow members of the subreddit community. We finally obtain 2174 samples, and annotate textual cues in 465 ($\approx$25%) positively labeled samples for TLC (*trigger*: 3988 words; *LoST indicators*: 6514 words; *consequences*: 787 words).

Tackling the intricate and highly subjective task of identifying low self-esteem in textual content can inadvertently lead to mistakes when relying solely on *naive* judgment. To address this challenge, our approach encompassed forming a collaborative unit comprising a clinical psychologist (tasked with discerning the nuanced psychological undertones within the text) and a social NLP expert (responsible for meticulous text annotations conducive for advanced AI models). To harmoniously blend insights from clinical psychology with the NLP realm, domain experts propose granular guidelines that categorize low self-esteem as an underlying psychological concept marked by self-doubt, feelings of worthlessness, and a discernible absence of confidence. Aiding the annotation chore, our domain experts contrived a structured annotation blueprint anchored on clinical questionnaires and two pivotal research queries: (i) “RQ1: *Is the text posted in the subreddits related to depression and suicidal ideation shows the sign of low self-esteem through direct text-spans ?*”, and (ii) “RQ2: *To what extent should annotators delve into the text to identifying text-spans for low self-esteem?*”.

Annotation Scheme When an individual experiences low self-esteem within the context of a mental health condition, healthcare professionals usually assign diagnostic codes that correspond to the specific mental health disorder contributing to or causing the low self-esteem. We define our annotation task as the identification of text-spans that encapsulate the essence of low self-esteem.

Given the inherently subjective and inference-driven nature of textual data, our research aims to develop and disseminate comprehensive annotation guidelines for discerning clinical concepts within text and subsequently translating them into diagnostic codes. It's worth noting that within the International Classification of Diseases, 10th Revision (ICD-10), there isn't a dedicated code specifically designated for “low self-esteem” as a primary diagnosis. The ICD-10 primarily focuses on classifying medical conditions and diseases, with mental health conditions, including issues related to self-esteem, typically categorized within broader classifications like mood disorders or neurotic disorders. Some of the major ICD codes are: F32 - Major depressive

S.No.	Text	Label
T1	... now i was diagnosed with mental illness and he is disappointed because he is afraid I will turn like my mother. Yesterday at night I went to the bathroom which is next to my parents bedroom and he said to my mom about her but then he also included me ”**i am disgusted by fat**← (`trigger`), its **disgusted, filthy and dirty**← (`trigger`) i don't even know how to express it to my daughter”. It just **broke me**← (`trigger`), broke me to pieces. *I doubt he even loves me*← (`LoST`), i have seen him looking at me with **disappointment**← (`trigger`) when i eat something it's just makes me *hate my self more*← (`LoST`) and make me WANT TO KILL MYSELF← (`consequences`)...	Presence
T2	I'm *boring*← (`LoST`). I'm *not good*← (`LoST`) at socializing and I'm very *awkward*← (`LoST`). I'm *replaceable*← (`LoST`). **** I WANT TO DIE← (`consequences`) so much right now.	Presence
T3	It's night and as usual that's when all these **horrible thoughts come to my head**← (`trigger`). I know I'm **not a terrible person** but I haven't exactly had the best year so I feel kind of overwhelmed right now. I really need to talk to someone ASAP I DON'T FEEL GOOD← (`consequences`).	Absence
T4	So, I have been thinking of HARMING MYSELF← (`consequences`) lately even though it's been a while since I've done it.	Absence

Table 1: Samples of the dataset. We label a given text along with three textual cues (i) triggering (bold text), (ii) LoST indicators (italicized text) and (iii) consequences (capitalized text).

disorder, Dysthymic disorder (Persistent depressive disorder), Panic disorder (with or without agoraphobia), Anxiety disorder, Obsessive-compulsive disorder, Reaction to severe stress, and adjustment disorders, Dissociative and conversion disorders, Somatoform disorders (which include somatic symptom disorders).

Our experimentation endeavors to establish the groundwork for automating the conversion process, enabling the transformation of text, whether it originates from user-generated content or clinical notes, into relevant diagnostic codes. We further utilize standard clinical questionnaires (SCQ) to frame annotation scheme to supervise the correct annotations. Within this framework, a collaboration between a clinical psychologist and a social NLP researcher has resulted in the creation of annotation guidelines based on structured clinical questions (SCQs). These SCQs encompass a collection of questions rooted in psychological principles, which are essential for extracting precise and pertinent information. Furthermore, these questions employ assessment tools to collect data and appraise a variety of psychological constructs. The foundation for our dataset creation lies in the SCQs derived from three primary clinical surveys: (i) Rosenberg's Self-esteem Scale (RSS) (Rosenberg 1965), (ii) Coopersmith Self-Esteem Inventory (CSEI) (Potard 2017), and (iii) Interpersonal Needs Questionnaire (INQ-18) (Mitchell et al. 2020) (see Figure 1). We take annotations to identify three types of text-spans in a given text: Triggering, LoST, and Consequences (TLC). We define TLC as follows:

1. **Triggering**: The reason behind low-self esteem is a triggering component that further enhances differentiation between the *low self-esteem* and an *event that may incur low self-esteem* in a person. We instruct the annotators to identify text-spans causing mental disturbance.
2. **LoST**: The text spans that indicates one of the 10 predefined annotation principles in the first-person context, should be identified as low self-esteem. For example, there is a substantial difference between `my friends think I am not funny` and `I am not funny`. Although the text-span `I am not funny` is present in both these statements, however, the former one should be marked as absence of low self-esteem due to public opinion which may be a seed to implant the prospective low self-esteem. However, the cross-sectional study cannot reveal users' perception about public opinion and hence, we do not make any assumptions.
3. **Consequences**: There is a substantial difference between text indicators of final state of mental disturbance (self-harm and suicidal ideation), and low self-esteem. We instruct the annotators to identify text-spans casting signs of severe consequences due to mental disturbance such as *'feels like the end is near', 'feeling trapped'*.

Our experts developed the annotation guidelines for identifying textual cues team leveraging SCQs. Three postgraduate students carry out the manual annotations after successfully completing experts-driven training sessions. Annotations were carried out using the expert-driven annotation scheme, designed to ensure **consistency** and **synchronization** during annotations.

Annotators were made to sit together and annotate the text spans for TLC, resulting in *one group-annotation* to facilitate coherent annotations. This annotation task was followed by experts' validation.

Inter-Annotator Agreement After experts' validation, we test the reliability of their judgement for all the 465 samples using *Fliess' Kappa inter-observer agreement* study (Guggenmoos-Holzmann 1996). We employ two experts to verify the annotations by marking them as either *acceptable* or *unacceptable*. To quantify the agreement, we calculate κ for *acceptance* and notice 67.52%, 71.92%, and 69.32% of agreement among annotators for *Trigger, LoST indicators, and Consequences*, respectively. We acknowledge that the lower value of inter-annotator agreement are a well-known problem in emotion-based subjective studies, where lower agreement scores are reported (Tsakalidis et al. 2018). The samples of our dataset in Table 1, exemplifies the concept of explainable low self-esteem detection. We observe that all three types of textual cues in TLC, are present in $T1$ but `trigger` is missing in $T2$. We illustrate that the presence of `consequences` does not ensure the presence of `LoST indicators` (see $T3$ in Table 1).

FAIR Principles The FAIR principle (Wilkinson et al. 2016), enhances the findability, accessibility, interoperability, and reusability of datasets to emphasize the actionable nature of machine-centric systems, which are becoming increasingly relied upon in facilitating future research endeavors. Our dataset contains the [*text*, *label*, *Trigger*, *LoST indicators*, and *Consequences*] for all the positively labeled data-points and later three labels (TLC) for the other negatively labeled data-points. The dataset is constructed in the comma-separated format[1]. We plan to expand and update our dataset with more data-points and more aspects of mental health in upcoming versions. By respecting the necessary safeguards, we aim to ensure the responsible use of this dataset while enabling advancements in understanding different aspects of mental health through computational linguistic approaches. In future, we plan to enhance the other aspects of our sister datasets such as explainable loneliness detection (Garg et al. 2023) on the same lines.

Inferences We acknowledge that only 33.12% of the positively labeled data-points contains all three types of textual cues among TLC. We find 465 $textual - cues$ for LoST indicators, overlapping with *Trigger* and *consequences* upto 85.16% and 40.64%, respectively. The natural composition of the dataset is $\approx\frac{3}{1}$ ratio for negative (0) to positive (1) label where positive (1) label indicates the presence of low self-esteem. The schema of our dataset is as follows:

> < ***Text*** *(string),* ***Label*** *(binary)>,*
> *// Text-classification problem*
>
> < ***Trigger*** *(list of string),*
> ***LoST indicators*** *(list of string),*
> ***Consequences*** *(list of string)>*
> *// Reliability analysis*

Furthermore, we examine $T3$ in Table 1, where the individual refers themselves as `'not a terrible person'`, indicating that the person does not possess any thoughts of low self-esteem. However, it's worth noting that words such as `'I'`, `'terrible'`, and `'person'` may not provide reliable predictions, resulting in negative perception of the user. To this end, we highlight the importance of considering semantic enhancements to develop more comprehensive and informative models.

Proposed Method

The idea behind reliability analysis is to identify text-spans that are aimed at detecting the presence of low self-esteem in user-penned content on social media. We apply *BERT*, a model tailored to detect text-spans in user-penned posts that potentially signify low self-esteem. The model systematically processes posts from a set P and classifies them accordingly.

Problem Formulation

Given a collection of posts, represented as $P = \{P_1, P_2, \ldots, P_n\}$, our model operates in two primary phases: *an attention mechanism* to discern the relevant text-spans and *a classification mechanism* to decide the overall sentiment of the post concerning self-esteem as shown in Figure 2.

Attention Mechanism The core of pre-trained language models is the attention mechanism that dynamically computes weights, or "attention scores", for different parts of the input, allowing the model to focus selectively on specific parts of the input, especially areas that may signal low self-esteem. Given a post P_i, each token in a post has an associated attention weight. Formally, the attention mechanism A for post P_i can be represented as:

$$A(P_i) = \{a_1, a_2, \ldots, a_m\} \tag{1}$$

Where:

- a_j stands for the attention weight assigned to the j^{th} token of post P_i.
- m signifies the total number of tokens within post P_i.

Classification Mechanism Once we compute the attention weights, the subsequent step involves classifying the text based on its content. The classification leans heavily on the attention scores, using them in combination with the embedded representations of the tokens to arrive at a binary decision. The function C represents this classification step. Given the attention weights from function A and the token embeddings, C produces an outcome indicating if the post manifests signs of low self-esteem:

$$C(A(P_i)) = \begin{cases} 1 & \text{if low self-esteem is detected} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

In essence, for every post P_i in our set P, the model follows a two-step process:

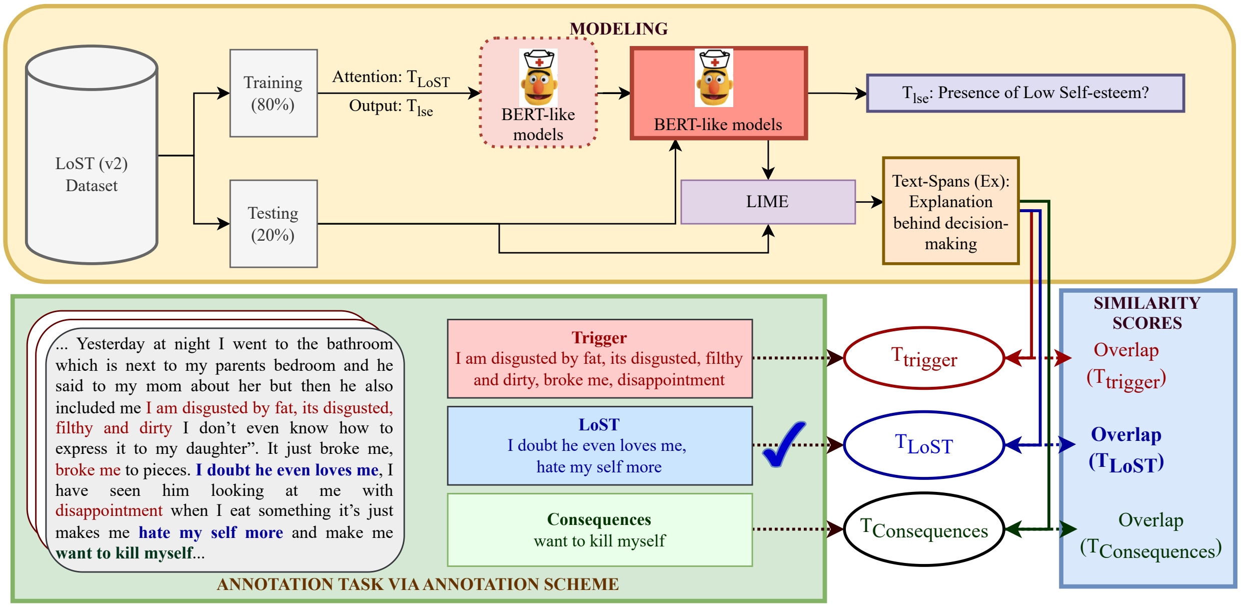
1. Compute attention weights using the function A.
2. Use these weights, along with the post's token embeddings, to classify the post with the function C. The result denotes the detected sentiment: 1 for low self-esteem and 0 otherwise.

Mathematical Notations

The core utility of attention mechanisms in NLP models, such as BERT, is to ascertain the significance of individual tokens or spans within larger sequences. For our objective of ***identifying indicators of low self-esteem*** in user-penned text, we aim to emphasize specific words or phrases that may suggest such clinical concepts.

1. *Token Embeddings:* Every token t within a post P_i is mapped to a dense vector, denoted as $E(t)$. This embedding captures the semantic nuance of the tokens, offering a representation in higher-dimensional space.

[1] The dataset shall be made available on request via signed agreement.



... Yesterday at night I went to the bathroom which is next to my parents bedroom and he said to my mom about her but then he also included me I am disgusted by fat, its disgusted, filthy and dirty I don't even know how to express it to my daughter". It just broke me, broke me to pieces. **I doubt he even loves me**, I have seen him looking at me with disappointment when I eat something it's just makes me **hate my self more** and make me **want to kill myself**...

Figure 2: Architecture of Reliability Analysis for Low self-esteem detection and classification in user-penned text.

2. *Transformations to Query, Key, and Value Vectors:* Each token is further transformed to generate its query, key, and value vectors. These are critical for determining the relative importance of a token in the context of the entire post.

$$Q(t) = E(t)W_Q$$
$$K(t) = E(t)W_K$$
$$V(t) = E(t)W_V$$

   Where: W_Q, W_K, and W_V are trainable weight matrices for the query, key, and value transformations respectively.

3. *Attention Score Computation:* For each token, an attention score s_j is derived to measure its relevance with respect to every other token, particularly in the context of detecting signs of low self-esteem:

$$s_j = Q(t_j) \cdot K(t)^T$$

4. *Normalization via Softmax:* The raw scores are normalized using the softmax function to make them lie between 0 and 1, ensuring they sum up to 1:

$$\alpha_j = \frac{\exp(s_j)}{\sum_{k=1}^{m} \exp(s_k)}$$

   Here, α_j is of utmost importance. Tokens with particularly elevated α_j values are ones that the model considers critical in signifying low self-esteem.

Upon computing the attention weights, our subsequent goal is to use this enriched data to ***classify*** the conceptual significance of the text, specifically, determining if the emphasized tokens suggest low self-esteem.

1. *Collated Post Representation:* Considering the varying attention weights for each token, we generate a consolidated representation for the post P_i:

$$R(P_i) = \sum_{j=1}^{m} \alpha_j A(t_j)$$

2. *Through the Classification Layer:* This combined vector is then pushed through a dense layer which outputs a probability score p. This score represents the likelihood that the post exhibits low self-esteem:

$$p = \sigma(R(P_i)W_C + b_C)$$

   Where:

   - W_C is the weight matrix associated with the classification layer.
   - b_C is the bias term.

3. *Binary Classification Based on Threshold:* Depending on the value of p, a binary classification decision is made:

$$C(A(P_i)) = \begin{cases} 1 & \text{if } p \geq 0.5 \text{ (indicative of low self-esteem)} \\ 0 & \text{otherwise} \end{cases}$$

For ***binary classification tasks***, one of the most commonly employed loss functions is the binary cross-entropy loss. Given the predicted probability p and the true label y (where $y = 1$ indicates low self-esteem and $y = 0$ otherwise), the binary cross-entropy loss, L, is defined as:

$$L(p, y) = -\left(y \log(p) + (1 - y) \log(1 - p)\right)$$

Where:

- p is the predicted probability of the post being indicative of low self-esteem.
- y is the ground truth label.

Reliability Analysis

In our research, we delve deeply into the specific text regions, commonly referred to as "text-spans", that the BERT model directs its attention towards, in an attempt to identify indicators of low self-esteem within written content. As per our established methodology, each piece of text T is meticulously annotated into several categories:

$$T = \{T_{lse}, T_{trigger}, T_{LoST}, T_{consequences}\} \tag{3}$$

where:

- T_{lse} denotes the presence or absence of low self-esteem.
- $T_{trigger}$ represents textual cues or triggers leading to potential mental disturbances.
- T_{LoST} indicates the textual span highlighting the author's perceived low self-esteem.
- $T_{consequences}$ captures the aftermath or resulting mental state from the disturbances.

Given the model's attention mechanism, for any token t_i in T, the attention weight is denoted as $A(t_i)$. Our hypothesis posits that the model's attention, when detecting low self-esteem, is primarily distributed among the tokens related to the three categories of trigger, LoST, and consequences. Mathematically, the attention distribution for each category is:

$$A_{category} = \sum_{t_i \in T_{category}} A(t_i) + \epsilon \tag{4}$$

where ϵ is the attention given to the words other than the ones in TLC category. Ideally, the attention weights should be skewed towards the LoST category. Therefore, an optimal model's attention distribution would satisfy:

$$A_{LoST} > A_{trigger} \tag{5}$$

$$A_{LoST} > A_{consequences} \tag{6}$$

To evaluate the model's accuracy in focusing on the correct text-spans, we utilize an *exact match algorithm*. For each category $category$ in the text T, the overlap with the model's explanations Ex is computed as shown in Equation 7:

$$Overlap_{category} = \frac{|Ex \cap T_{category}|}{|T_{category}|} \tag{7}$$

where $category$ belongs to a set containing all three categories: $\{T_{trigger}, T_{LoST}, T_{consequences}\}$. The overlap gives a percentage indication of how well the model's attention or explanation aligns with the pre-annotated TLC text-spans.

Experiments and Evaluation

Classifiers as Baselines

We on several models originating from the BERT architecture, each bringing its unique features to the table, specifically for concept extraction and classification.

- **BERT** is a transformer-based architecture that captures contextual information from both left and right sides of a token in any input text (Devlin et al. 2018). This architecture has set a new standard for a range of NLP tasks. The bi-directionality of BERT ensures that each word is analyzed in its surrounding context, making it potent for discerning subtle cues indicative of low self-esteem, which can be context-dependent.
- **ALBERT** is a variant of BERT optimized for faster training and less memory consumption without compromising on performance (Lan et al. 2019). ALBERT's parameter-reduction techniques ensure that we maintain the power of BERT while achieving faster model training, which is beneficial when working with large and nuanced datasets, such as those involving human emotions and states.
- **DistilBERT** is a distilled version of BERT, retaining most of its performance capabilities but being 40% smaller and 60% faster (Sanh et al. 2019). Given the need for real-time or faster processing in social media content analysis, DistilBERT's speed and size advantages make it an attractive choice for extracting low self-esteem indicators without significant lag.
- **DeBERTa** improves upon BERT by disentangling the inter-token semantic relations with absolute positional encoding (He et al. 2020). The disentangled attention mechanism can be pivotal in decoding intricate user emotions and sentiments, ensuring that the model is sensitive to both the semantics and position of tokens when identifying low self-esteem cues.
- **ClinicalBERT**, as its name suggests, is fine-tuned on clinical narratives or medical literature, making it adept at understanding medical terminologies and contexts (Huang, Altosaar, and Ranganath 2019). When considering low self-esteem within a clinical or medical paradigm, ClinicalBERT's expertise can ensure that medical or clinical references related to self-esteem are accurately captured and classified.
- Being optimized for psychological contexts, **PsychBERT** is inherently more sensitive and attuned to detecting subtle emotional indicators, such as those found in content penned by individuals with low self-esteem (Vajre et al. 2021).
- Addressing the wider spectrum of mental health, **MentalBERT** offers a holistic approach to detect low self-esteem, considering it alongside other potential mental health indicators (Ji et al. 2022).

Experimental Setup We implement the existing classifiers to test the accuracy and robustness of the models for identifying textual cues projecting low self-esteem. We split our dataset into a ratio of 80:20, with 80% of the data allocated for training samples and 20% reserved for testing purposes. The randomised distribution of training to testing data contains 376 positive samples out of 1,739 for training data, and 89 positive samples out of 435 samples. We use the validation set from training samples to fine-tune model hyperparameters and assess the performance during the training process. Among the classifiers, we consider four well-established pre-trained language models (PLMs) (BERT, ALBERT, DistilBERT, and DeBERTa), while the remaining three are domain-specific PLMs having strong association with mental health domain. We use the standard PLMs available on huggingface and default attention mechanism to perform experiments with existing classifiers. We set the learning rate (lr) as $2e-5$ for a batch size of 8, weight decay as 0.01, warmup steps of 100, and logging steps of 100 for fine-tuning PLMs. Furthermore, we obtain a new testing dataset of 200 samples from out-of-distribution (OOD) dataset, Dreddit, a publicly available data (Turcan and McKeown 2019), originally constructed to classify depression and suicide risk. We annotate first 200 instances of Dreddit dataset for reliability analysis using the same annotation scheme. We use the Google colab pro environment for experiments and evaluation to access faster GPUs like the Tesla P100, and occasionally the Tesla T4 and Tesla V100. This was beneficial for running compute-intensive tasks.

Evaluation metrics In this research paper, we employ *precision, recall, F-score, accuracy,* and *Matthew's correlation coefficient (MCC) (Boughorbel, Jarray, and El-Anbari 2017)* as evaluation metrics for identifying Reddit posts casting low self-esteem. MCC takes into account true and false positives and negatives and is generally regarded as a balanced measure, even when classes are of very different sizes. Additionally, we use LIME (Ribeiro, Singh, and Guestrin 2016) that generates faithful local explanations, to find system-level explanations and compare the resulting explanations with the ground-truth textual cues for (i) triggers, (ii) LoST indicators, and (iii) consequences, using two text similarity matching mechanisms (Yang et al. 2018): (i) Recall Oriented Understudy for Gisting Evaluation (ROUGE) scores, and (ii) BiLingual Evaluation Understudy (BLEU) scores. ROUGE provides several measures to determine the quality of a summary by comparing it with other (reference) summaries. BLEU compares the n-grams in the machine-generated translations to those in the reference translations and returns a score between 0 and 1, where 1 means the generated translation matches the reference perfectly.

Evaluation of Classifiers Table 2 compares the performance of all the existing classifiers. The domain-specific PLMs, MentalBERT (Ji et al. 2022) shows the best performance with the highest scores across almost all evaluation metrics (with second best results for Precision after PsychBERT (Vajre et al. 2021)), suggesting more accurate classification. This happens probably due to the domain-specific pre-training of the model on mental health-related posts collected from Reddit. Among the PLMs, BERT demonstrates the highest accuracy (0.8552) followed by AlBERT (0.8529). DistilBERT has the lowest scores for all the evaluation metrics, minimally contributing towards classifying texts with low self-esteem. ClinicalBERT (Huang, Altosaar, and Ranganath 2019) has the lowest F1-score (0.5914) and MCC (0.4808) because it is trained on clinical notes, resulting in indifferent nature of the text which is not suitable for our task. As such, the MentalBERT outperforms all other models followed by the BERT model, especially in terms of F-score, Accuracy and MCC score, suggesting clear interpretation of the classifier, overall correctness of the classifier, and efficiency of the model in-case of imbalanced dataset, respectively.

Model	Present			Acc.	MCC
	P	R	F		
BERT	0.6392	0.6889	**0.6631**	**0.8552**	**0.5717**
AlBERT	**0.6413**	0.6556	0.6483	0.8529	0.5554
DistilBERT	0.5978	0.6111	0.6044	0.8345	0.4998
DeBERTa	0.6095	**0.7111**	0.6564	0.8459	0.5606
MentalBERT	0.6500	**0.7222**	**0.6842**	**0.8621**	**0.5976**
ClinicalBERT	0.5729	0.6111	0.5914	0.8253	0.4808
PsychBERT	**0.6528**	0.5222	0.5802	0.8437	0.4902

Table 2: Detecting Low Self-Esteem. Comparison of the existing classifiers with Precision (P), Recall (R), F1-score (F1), Accuracy and MCC score, are averaged over 10-fold cross validation. Present: Presence of Low Self-Esteem.

Evaluation of Explanations We further evaluate the text-spans focused by attention mechanism to detect low self-esteem. Table 3 offers an in-depth look into the system-level explainability of various NLP models. Through this analysis, we aim to uncover how each classifier deciphers textual cues within a corpus. Using the LIME approach, we extract explanations and gauge their alignment with the three textual cues: *Trigger, LoST indicators*, and *Consequences* by leveraging two similarity metrics: ROUGE and BLEU. A noticeable trend is the superior performance of classifiers that closely align their system-level explanations with the "LoST indicators." This implies a fundamental shift in understanding classifier behavior: models that prioritize recognizing indicators related to LoST generally outdo those emphasizing "Trigger" or "Consequences". BERT and MentalBERT are exemplary in this regard, leading among the pre-trained Language Models (PLMs) and domain-specific pre-trained language models, respectively. However, models like AlBERT and PsychBERT which exhibited top scores in *Trigger*, and DeBERTa and ClinicalBERT which did well in *Consequences*, underscore a potential inefficiency. Ideally, these models should prioritize *LoST indicators* as key discriminative cues, but instead, they seem to focus on possibly less relevant textual signals. Further granulating our observations, we note that for *LoST indicators*, scores are consistently higher for True Positives (TP) than True Negatives (TN). This indicates a robust attention mechanism, particularly for TP instances, reaffirming that the models discern crucial cues correctly when classifying true instances.

While the classifiers show a promising tilt towards *LoST*, there's still significant alignment with *Triggers* and *Conse-*

Model	Evaluation	**TRIGGER (↓)**			**LOST INDICATORS (↑)**			**CONSEQUENCES (↓)**		
		ATS	**TP**	**TN**	**ATS**	**TP**	**TN**	**ATS**	**TP**	**TN**
BERT	ROUGE	0.0756	0.0705	0.0868	**0.2773**	***0.3128***	0.1985	0.0421	0.0442	0.0375
	BLEU	0.0519	0.0446	0.0681	**0.1656**	***0.1760***	0.1426	0.0315	***0.0349***	0.0241
ALBERT	ROUGE	**0.0987**	***0.0973***	**0.1014**	0.2201	0.2424	0.1777	0.0359	0.0427	0.0229
	BLEU	**0.0657**	***0.0595***	**0.0774**	0.1363	0.1367	0.1356	0.0271	0.0318	0.0179
DistilBERT	ROUGE	0.0826	0.0837	0.0809	0.2471	0.2738	0.2051	0.0351	0.0273	0.0474
	BLEU	0.0569	0.0564	0.0579	0.1451	0.1412	0.1513	0.0260	0.0219	0.0324
DeBERTa	ROUGE	0.0687	0.0645	0.0790	0.2461	0.2570	**0.2193**	**0.0468**	***0.0465***	**0.0476**
	BLEU	0.0459	0.0389	0.0634	0.1586	0.1513	**0.1724**	**0.0346**	0.0341	**0.0359**
ClinicalBERT	ROUGE	0.0728	0.0729	0.0726	0.2045	0.2304	0.1639	**0.0369**	***0.0359***	0.0387
	BLEU	0.0463	***0.0451***	0.0485	0.1260	0.1272	0.1241	**0.0286**	***0.0288***	0.0284
PsychBERT	ROUGE	**0.0946**	0.0729	**0.1272**	0.2310	0.2304	0.1949	0.0307	***0.0359***	**0.0415**
	BLEU	**0.0648**	***0.0451***	**0.0860**	0.1336	0.1272	0.1414	0.0226	***0.0288***	**0.0325**
MentalBERT	ROUGE	0.0820	***0.0733***	0.1047	**0.2614**	***0.2821***	**0.2076**	0.0160	0.0126	0.0248
	BLEU	0.0464	0.0348	0.0768	**0.1544**	***0.1521***	**0.1604**	0.0107	0.0091	0.0147

Table 3: Comparison of the explanations obtained by LIME with three different types of textual cues annotated as `Trigger, LoST indicators, and consequences` using ROUGE scores and BLEU scores. ([Black, italicize, underlined] + bold) represents the highest values of [ATS: All Test Samples, TP: True Positives, TN: True Negatives], respectively. Higher the value of LoST indicators and lower (↓) the values with Trigger and consequences, better is the classifier.

quences. This can make explanations unclear, causing confusion and leading to potential misinterpretations. The findings underline a significant challenge: creating models that are better calibrated towards specific indicators, like *LoST*, instead of general cues like *Triggers* or *Consequences*. The latter may be more abundant or explicit in texts but may not always be the most crucial from a psychological perspective. Future research should explore mechanisms to prioritize *LoST indicators* during model training. This could involve novel attention mechanisms, penalization techniques during training to reduce emphasis on *Triggers* and *Consequences*, or even dataset augmentation to include more explicit *LoST indicators* instances. As NLP models delve deeper into psychological text analysis, their explainability becomes paramount. It's crucial not just to have accurate models but also ones that can clearly and intuitively explain their decision-making rationale, especially in sensitive areas like psychological analysis. Domain-specific PLMs like MentalBERT suggest the utility of fine-tuning on specialized domain-specific corpora. Further explorations could involve curating more granular psychological datasets and refining models on them to better capture nuanced indicators.

Out of Distribution Analysis We test the models on OOD dataset. We carry out the annotation task using the same annotation scheme on the 200 samples of existing dataset and examine classification method in Table **??**. From the general-purpose models, BERT emerged as top performer in terms of F1-score, achieving 0.6667. The balance between precision and recall for BERT demonstrates its capacity to identify low self-esteem indications reliably without raising too many false alarms. DeBERTa's numbers echo this sentiment, with even a slight improvement in accuracy. MentalBERT, as expected from a domain-focused model, exhibited remarkable precision at 0.9231, suggesting that when it flags a text-span as indicative of low self-esteem, it's likely correct. However, its recall sits at 0.4800, hinting that it might miss out on some relevant instances. This reinforces the idea that while domain-specific models might be more accurate in their detections, ensuring reliability in terms of comprehensiveness remains a challenge. Metrics like the MCC play a pivotal role in understanding a model's reliability. For instance, BERT's MCC score of 0.6295 indicates a good quality binary classification, while models with lower MCC scores, like ClinicalBERT at 0.4280, might not be as reliable in distinguishing between the positive and negative classes. The results underscore the fact that while several models can detect low self-esteem indications with decent accuracy, achieving reliable and consistent results is a nuanced challenge. It's not just about flagging potential indicators; it's about ensuring that these detections are genuine, consistent, and that potential indicators aren't overlooked. While BERT present promise, there's room for improvement, especially in the balance between precision and recall. As the stakes involve individuals' psychological well-being, a miss or false alarm isn't trivial. Thus, the quest for a model that can reliably analyze and flag text-spans indicative of low self-esteem on social media platforms, given the ever-evolving nature of online discourse, remains an ongoing challenge.

An OOD dataset provides an essential playground to test the robustness of classifiers. It's crucial to see how well models generalize and what kind of explanations they provide when confronted with data that is not part of the original dataset. Table **??** captures how the widely acknowledged explanation framework, LIME, explains the models' decisions using ROUGE and BLEU scores. ALBERT and DistilBERT's explanations, as indicated by its scores, appear to be somewhat skewed towards *Trigger*. Thus, ALBERT and DistilBERT might sometimes conflate triggers or consequences with actual indicators of low self-esteem. DeBERTa's scores are interesting. Its explanations tend to blur the lines between triggers, and indicators. The difference in scores is

Model	Present			Acc.	MCC
	P	R	F		
BERT	0.750	0.600	**0.666**	**0.924**	**0.629**
AlBERT	0.846	0.440	0.578	0.919	0.574
DistilBERT	0.750	0.360	0.486	0.904	0.477
DeBERTa	0.823	0.560	***0.632***	***0.919***	0.613
MentalBERT	0.923	0.480	**0.632**	**0.929**	**0.636**
ClinicalBERT	0.777	0.280	0.411	0.899	0.428
PsychBERT	0.684	0.520	0.590	0.909	0.547

Table 4: Identification of Low Self-Esteem with LoST.v2 dataset. Comparison of the existing classifiers with Precision (P), Recall (R), F1-score (F1), Accuracy and MCC score, are averaged over 10-fold cross validation. Present: Presence of Low Self-Esteem.

not significant, implying that DeBERTa might have difficulty distinctly recognizing each type of textual cue. ClinicalBERT's scores reiterate the challenge of distinctly identifying textual cues. While its performance on recognizing LoST indicators is commendable, it seems to also give substantial weight to consequences, as seen from its relatively high BLEU score. PsychBERT stands out, with its ROUGE score for LoST indicators (0.2310) being notably higher than the scores for triggers and consequences. This suggests that PsychBERT's explanations are well-aligned with recognizing actual indicators of low self-esteem. Lastly, MentalBERT, another domain-specific model, also shows a notable distinction in scores, particularly with a higher emphasis on LoST indicators, which is encouraging. It indicates that MentalBERT, in most cases, recognizes and emphasizes the actual indicators over triggers and consequences. However, the performance of MentalBERT is compromised as compare to BERT due to 2.5 times more focus on *Trigger*.

Discussion

In the realm of NLP, we benchmark traditional classifiers to evaluate the performance. However, to enhance the scope of our research, we focus our attention on two paramount models: *BERT* and *MentalBERT*.

Performance. With reference to the *ROUGE* score, a widely accepted standard for evaluating text-based tasks, our results were elucidating. The ratio of *[LoST indicators]* with *[Trigger + consequences]* together yield intriguing results. *BERT* registered a ratio of approximately 2.354, while *MentalBERT* showcased a superior ratio of approximately 2.667. This distinction in performance is further mirrored when considering the *BLEU* scores. BERT posted a ratio of roughly 1.987, while MentalBERT notably exceeded this with a ratio close to 2.704. Drawing from these numerical evaluations, it is evident that, based on our empirical data and selected metrics, *MentalBERT* emerges as the more robust model in comparison to the conventional *BERT* framework. However, *MentalBERT* compromise the performance on OOD data, suggesting *BERT* as more reliable model than *MentalBERT*. We plan to examine the robustness and trustworthiness of two models in the future work.

Model	Eval.	T	L	C
BERT	ROUGE	0.0449	*0.1764*	0.0533
	BLEU	0.0225	*0.1328*	0.0432
ALBERT	ROUGE	0.1525	0.1857	0.0259
	BLEU	0.1142	0.1468	0.0166
DistilBERT	ROUGE	0.0569	0.1291	0.0266
	BLEU	0.0138	0.0993	0.0200
DeBERTa	ROUGE	0.0979	0.1094	0.0159
	BLEU	0.0638	0.0739	0.0100
ClinicalBERT	ROUGE	0.0607	0.0968	0.0533
	BLEU	0.0279	0.0536	0.0500
PsychBERT	ROUGE	0.1444	*0.2310*	0.0352
	BLEU	0.1018	*0.1843*	0.0260
MentalBERT	ROUGE	0.1054	*0.2228*	0.0214
	BLEU	0.0633	*0.1897*	0.0133

Table 5: Observations with the Out-of-distribution (OOD) dataset. Comparison of the explanations obtained by LIME with three different types of textual cues annotated as `Trigger`, `LoST indicators`, `and consequences` using ROUGE scores and BLEU scores. (italicize and underline) represents the highest values of LoST in True Positives and comparable values of LoST and Trigger, respectively.

Attention Mechanism Error Analysis. Notably, the inherent attention mechanism in these classifiers displayed tendencies to stray from concentrating on the *LoST indicators*. This diversion is significant, constituting between 30% and 50% of the textual cues vital for crafting system-level explanations. To address these observed limitations and to steer the discipline towards a more promising direction, we have curated the our dataset for reliability analysis and OOD testing.

Introducing a New Dataset. Our dataset is emblematic of a broader vision: urging the scholarly community to prioritize models accentuating trustworthiness, safety, and, most importantly, reliability. While accuracy indisputably remains a pivotal metric, our research underscores the pressing need to transcend the enticements of mere accuracy offered by opaque "black-box" NLP models and to champion transparency and intelligibility.

Significance of Annotation Consistency. The integrity of any open-source dataset largely hinges on the consistency and reliability of its annotations. When working with datasets, especially in the realm of NLP, ensuring that each data point is annotated with precision and uniformity is paramount. This not only serves to validate the authenticity and reliability of the dataset but also offers researchers and model designers a clear comprehension of the underlying structure and nature of the data. As such, achieving consistency in annotations becomes an indispensable step for avoiding biases, ensuring replicability, and ultimately obtaining robust results across various applications and analyses.

Role of Textual Span Categorization. The systematic categorization and demarcation of textual spans into three distinct types plays a pivotal role in guiding models towards the intricate task of identifying instances of low self-esteem in texts. By providing a clear framework for these spans, models are endowed with the necessary guidance to discern and recognize the subtle nuances and patterns indicative of low self-esteem.

Ethics and Broader Impact. Our dedication lies in upholding ethical principles to safeguard user privacy and anonymity (Henderson et al. 2018). To prevent any misuse, the examples presented in this paper are modified through obfuscation, and paraphrasing. In order to uphold the ethical principles of privacy, safety, and accountability, we have abstained from disclosing any metadata in the public domain. Due to the subjective nature of our task, there may be some inherent biases in our annotations (Zirikly and Dredze 2022). As we consider explainability as the decision-making parameter, we encourage the enhancement of classifier's attention mechanism in the near future. We design our dataset to facilitate the automated and reliable annotations in identifying various aspects of mental disturbance within a given text (Meyer et al. 2022; Wang et al. 2021). The practical application of this NLP-centered task is the pre-screening of social media users during in-person session of mental health triaging, clinical diagnostic interviewing and motivational interviewing (Daws 2020; Westra, Aviram, and Doell 2011). Moreover, this task elicits both risk and resilience features when monitoring cognitive decline and severe mental disorders. Another practical considerations is its applicability to problems with work-life balance, abusive relationships and impact of job-layoffs during economic recession (Heron, Eisma, and Browne 2022; Howard et al. 2022). We acknowledge the need of its licensed use by clinicians, practitioners and other stakeholders to avoid any potential misuse or societal impact of our work.

Limitations. While BERT-based architectures are known for their accuracy, they remain largely "black-box" in nature. Despite using evaluation metrics to understand their performance, the underlying reasons for specific classifications are not always transparent, which can be crucial when dealing with sensitive topics like mental health. The performance of models like BERT and its variants largely depends on the quality and quantity of training data. If the dataset does not comprehensively represent the diversity of expressions of low self-esteem across various demographic and sociocultural groups, the models' generalization capabilities may be limited. Models such as PsychBERT, MentalBERT, and ClinicalBERT, while fine-tuned for specific psychological task in mental health domain, might still miss subtle textual-cues. While it's essential to assess the models' performance in unseen data, such evaluations can sometimes lead to results that don't accurately reflect real-world applicability. Thus, OOD dataset evaluation presents its challenges.

Conclusion

As the NLP research community progresses towards developing system-level explainable classifiers, our corpus will play a crucial role in advancing the field of information retrieval in the near future. Our task of constructing an advanced corpus, reveals that the classifiers have shown an increased focus on textual cues that emphasize low self-esteem in Reddit posts and emphasise the pressing need of reliability and robust models for healthcare utilization. The annotated explanations in dataset show a closer alignment with the `LoST indicators`, although significant similarities still exist with the textual cues indicating triggers and consequences. We establish the BERT model trained over the T_{LoST} text-spans for attention and T_{lse} for classification mechanism, as baseline. We further test its reliability using LIME for extracting explanations or focused text-spans by PLM's and testing over OOD dataset to examine the robustness of the classifiers. In future, it would be interesting to develop efficient models that redirects the attention of NLP models from *triggers* and *consequences* towards *LoST indicators* by infusing external knowledge such as domain-specific knowledge graph and commonsense knowledge.

Acknowledgements

This study was supported by NIH (National Institutes of Health) R01 AG068007.

References

Acarturk, C.; Smit, F.; De Graaf, R.; Van Straten, A.; Ten Have, M.; and Cuijpers, P. 2009. Incidence of social phobia and identification of its risk indicators: a model for prevention. *Acta Psychiatrica Scandinavica*, 119(1): 62–70.

Boughorbel, S.; Jarray, F.; and El-Anbari, M. 2017. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PloS one*, 12(6): e0177678.

Cella, S.; Cipriano, A.; Aprea, C.; Milano, W.; Carizzone, F.; and Cotrufo, P. 2022. Non-suicidal self-injury in eating disorders: Prevalence, characteristics, DSM-5 proposed diagnostic criteria, and correlates. *Journal of Affective Disorders Reports*, 7: 100292.

Choi, K. H.; Wang, S.-M.; Yeon, B.; Suh, S.-Y.; Oh, Y.; Lee, H.-K.; Kweon, Y.-S.; Lee, C. T.; and Lee, K.-U. 2013. Risk and protective factors predicting multiple suicide attempts. *Psychiatry research*, 210(3): 957–961.

Collaborators, G. . M. D.; et al. 2022. Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *The Lancet Psychiatry*, 9(2): 137–150.

Cox, R. 1987. The rich harvest of Abraham Maslow. *Motivation and personality*, 245–271.

Daws, R. 2020. Babylon Health lashes out at doctor who raised AI chatbot safety concerns. https://www.artificialintelligence-news.com/2020/02/26/babylon-health-doctor-ai-chatbot-safety-concerns/.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Garg, M.; Saxena, C.; Samanta, D.; and Dorr, B. J. 2023. LonXplain: Lonesomeness as a Consequence of Mental Disturbance in Reddit Posts. *arXiv preprint arXiv:2305.18736*.

Guggenmoos-Holzmann, I. 1996. The meaning of kappa: probabilistic concepts of reliability and validity revisited. *Journal of clinical epidemiology*, 49(7): 775–782.

He, P.; Liu, X.; Gao, J.; and Chen, W. 2020. DEBERTA: DECODING-ENHANCED BERT WITH DISENTANGLED ATTENTION. In *International Conference on Learning Representations*.

Henderson et al., P. 2018. Ethical challenges in data-driven dialogue systems.

Heron, R. L.; Eisma, M.; and Browne, K. 2022. Why do female domestic violence victims remain in or leave abusive relationships? A qualitative study. *Journal of Aggression, Maltreatment & Trauma*, 31(5): 677–694.

Howard, M. C.; Follmer, K. B.; Smith, M. B.; Tucker, R. P.; and Van Zandt, E. C. 2022. Work and suicide: An interdisciplinary systematic literature review. *Journal of Organizational Behavior*, 43(2): 260–285.

Huang, K.; Altosaar, J.; and Ranganath, R. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

Ji, S.; Zhang, T.; Ansari, L.; Fu, J.; Tiwari, P.; and Cambria, E. 2022. MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 7184–7190.

Kolubinski, D. C.; Frings, D.; Nikčević, A. V.; Lawrence, J. A.; and Spada, M. M. 2018. A systematic review and meta-analysis of CBT interventions based on the Fennell model of low self-esteem. *Psychiatry research*, 267: 296–305.

Korkmaz, H.; Korkmaz, S.; and Çakar, M. 2019. Suicide risk in chronic heart failure patients and its association with depression, hopelessness and self esteem. *Journal of clinical neuroscience*, 68: 51–54.

Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; and Soricut, R. 2019. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. arXiv:1909.11942.

Meyer, S.; Elsweiler, D.; Ludwig, B.; Fernandez-Pichel, M.; and Losada, D. E. 2022. Do We Still Need Human Assessors? Prompt-Based GPT-3 User Simulation in Conversational AI. In *Proceedings of the 4th Conference on Conversational User Interfaces*, 1–6.

Mitchell, S. M.; Brown, S. L.; Roush, J. F.; Tucker, R. P.; Cukrowicz, K. C.; and Joiner, T. E. 2020. The Interpersonal Needs Questionnaire: Statistical considerations for improved clinical application. *Assessment*, 27(3): 621–637.

Potard, C. 2017. Self-esteem inventory (Coopersmith).

Raza, S.; and Schwartz, B. 2023. Constructing a disease database and using natural language processing to capture and standardize free text clinical information. *Scientific Reports*, 13(1): 8591.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144.

Rosenberg, M. 1965. Rosenberg self-esteem scale. *Journal of Religion and Health*.

Rouault, M.; Will, G.-J.; Fleming, S. M.; and Dolan, R. J. 2022. Low self-esteem and the formation of global self-performance estimates in emerging adulthood. *Translational Psychiatry*, 12(1): 272.

Sanh, V.; Debut, L.; Chaumond, J.; and Wolf, T. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Sheth, A.; Gaur, M.; Roy, K.; and Faldu, K. 2021. Knowledge-intensive language understanding for explainable AI. *IEEE Internet Computing*, 25(5): 19–24.

Tsakalidis, A.; Papadopoulos, S.; Voskaki, R.; Ioannidou, K.; Boididou, C.; Cristea, A. I.; Liakata, M.; and Kompatsiaris, Y. 2018. Building and evaluating resources for sentiment analysis in the Greek language. *Language resources and evaluation*, 52(4): 1021–1044.

Turcan, E.; and McKeown, K. 2019. Dreaddit: A Reddit dataset for stress analysis in social media. *arXiv preprint arXiv:1911.00133*.

Vajre, V.; Naylor, M.; Kamath, U.; and Shehu, A. 2021. PsychBERT: a mental health language model for social media mental health behavioral analysis. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 1077–1082. IEEE.

Walker, E. R.; McGee, R. E.; and Druss, B. G. 2015. Mortality in mental disorders and global disease burden implications: a systematic review and meta-analysis. *JAMA psychiatry*, 72(4): 334–341.

Wang, S.; Liu, Y.; Xu, Y.; Zhu, C.; and Zeng, M. 2021. Want To Reduce Labeling Cost? GPT-3 Can Help. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 4195–4205.

Watson, J.; and Nesdale, D. 2012. Rejection sensitivity, social withdrawal, and loneliness in young adults. *Journal of Applied Social Psychology*, 42(8): 1984–2005.

Westra, H. A.; Aviram, A.; and Doell, F. K. 2011. Extending motivational interviewing to the treatment of major mental health problems: current directions and evidence. *The Canadian Journal of Psychiatry*.

Wilkinson, M. D.; Dumontier, M.; Aalbersberg, I. J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L. B.; Bourne, P. E.; et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3(1): 1–9.

Yang, A.; Liu, K.; Liu, J.; Lyu, Y.; and Li, S. 2018. Adaptations of ROUGE and BLEU to Better Evaluate Machine Reading Comprehension Task. *ACL 2018*, 98.

Zirikly, A.; and Dredze, M. 2022. Explaining Models of Mental Health via Clinically Grounded Auxiliary Tasks. In *CLPsych*.

Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes, this research advance towards reliability analysis of NLP models for healthcare utilization.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes, we define the task as the first of its kind to detect appropriate textual-cues at aid in decision-making of low self-esteem (psychological concept) detection.
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? The data is acquired based on the subreddit, irrespective of the population-specific distribution.
   (e) Did you describe the limitations of your work? Yes,
   (f) Did you discuss any potential negative societal impacts of your work? Yes, in the ethical and broader impact.
   (g) Did you discuss any potential misuse of your work? Yes.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, mentioned in FAIR principles, and ethics and broader impact.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? Partially applicable. Yes.
   (b) Have you provided justifications for all theoretical results? Partially applicable. Yes.
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? Partially applicable. Yes.
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? No such reliability analysis for domain-specific NLP models have been quantified for mental health analysis.
   (e) Did you address potential biases or limitations in your theoretical framework? Yes.
   (f) Have you related your theoretical results to the existing literature in social science? Partially applicable. Yes, grounded in psychological theories.
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? Yes,
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? Yes, in corpus construction.
   (b) Did you include complete proofs of all theoretical results? Yes,
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? No, because the code of the dataset shall be available on acceptance. We have build models with the traditional classifiers. The dataset will be available on request via signed agreement to adhere to ethical guidelines.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? Partially, in error analysis.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? NA
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? Yes
   (b) Did you mention the license of the assets? NA
   (c) Did you include any new assets in the supplemental material or as a URL? Yes, the dataset will be available on request via signed agreement to adhere to ethical guidelines.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? The dataset is curated via social media platform.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes, in ethics and broader impact.
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see **?**)? Yes
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see **?**)? The dataset will be available on request via signed agreement to adhere to ethical guidelines.
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
   (a) Did you include the full text of instructions given to participants and screenshots? NA

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA

(d) Did you discuss how data is stored, shared, and deidentified? NA