The Devil Is in the Details: Enhancing LLMs for Self-Harm Detection Through Intent Differentiation and Emoji Interpretation

Anonymous submission

Technical Appendices

Sub-reddits Used for Data Collection

Belos is the full list of sub-reddits considered for posts collection to build our SHINES dataset.

'mentalhealth', 'traumatoolbox', 'TrueOffMyChest', 'anxiety', 'BPD', 'depression', 'suicidewatch', 'mentalillness', 'selfharm', 'offmychest', 'vent', 'suicidalthinking', 'anxiety', 'operation', 'stress', 'competition', 'workPressure','sports', 'heavyHeart', 'mentalhealth', 'mentalillness', 'depression', 'politics', 'askatherapist', 'socialskills', 'BodyAcceptance', 'bodyneutrality', 'BodyNeutrality', 'Mindfulness', 'BipolarReddit', 'ADHD', 'bipolar', 'positivity', 'suicidewatch', 'suicidalthinking', 'officePolitics', 'parenting', 'selfinjurysupport', 'medication', 'characters', 'nostalgia', 'environment', 'instagram', 'relationships', 'panera', 'religion', 'selfhelp'.

Experimental Setup: Additional Details

We configured the learning rate to 4e-5 and employed a linear decay scheduler with a warm-up phase covering 10% of the total steps. A batch size of 16 provided a balance between training stability and speed. Training was conducted over 4 epochs, with validation loss closely monitored to prevent overfitting. The sequence length was set to 256 tokens, suitable for most social media posts. We used the AdamW optimizer with a weight decay of 0.01 to enhance generalization. A dropout rate of 0.2 was implemented to avoid overfitting, and gradient clipping at 1.0 ensured training stability. Model evaluation was performed after each epoch, with early stopping activated if no improvement was observed after 3 epochs. Warm-up steps were set to 500 to allow the model to adjust smoothly to the learning rate.

Metrics for Rationale Generation Evaluation

- **Relevance**
  - **Intuition**: This metric checks if the rationale includes all the key phrases related to casual mentions and serious intents.
  - **Functionality**:
    - Combines the casual mentions and serious intents into a single text.
    - Converts both the combined text and the rationale to lowercase.
    - Checks if all spans are present in the rationale.
- **Coherence**
  - **Intuition**: This metric evaluates the logical consistency and smoothness of the rationale in relation to the spans.
  - **Functionality**:
    - Combines the casual mentions and serious intents into a single text.
    - Uses TF-IDF Vectorizer to transform both the combined text and the rationale into vectors.
    - Computes cosine similarity between these vectors.
- **Readability**
  - **Intuition**: This metric measures how easy it is to read and understand the rationale.
  - **Functionality**:
    - Uses the Flesch-Kincaid grade level formula to compute the readability score of the rationale.
- **Semantic Similarity**
  - **Intuition**: This metric evaluates the conceptual similarity between the rationale and the spans.
  - **Functionality**:
    - Combines the casual mentions and serious intents into a single text.
    - Uses a pre-trained sentence embedding model (all-MiniLM-L6-v2)[1] to generate embeddings for both the combined text and the rationale.

[1] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

* Computes cosine similarity between these embeddings.

By employing these metrics, we can ensure that the generated rationales for self-harm detection are accurate, consistent, readable, and semantically aligned with the extracted spans.

Prompts

We present the prompts for the various baseline setups in Table 1. We present the prompts used for generating the synthetic data in Table 2

Type	Prompt
Zero-Shot	**## Instruction** You are tasked with evaluating social media posts to classify them as self-harm or non self-harm. Review the post below and provide your classification. Additionally, write a detailed rationale that explains your decision, taking into account the content and context of the post. **## Input** {post text} **## Output** **Classification**: [self-harm / non self-harm] **Rationale**: Describes the factors and considerations that guided the classification decision.
Few-Shot	**## Instruction** You are tasked with evaluating social media posts to classify them as self-harm or non self-harm. Review the following posts and classifications to understand the task. Then classify the new post and write a detailed rationale that explains your decision, taking into account the content and context of the post. **## Input** **Example 1**: **Post**: "I feel so worthless. Sometimes I think about hurting myself to make the pain stop." **Classification**: self-harm **Rationale**: The mention of "hurting myself" suggests a serious consideration of self-harm. **Example 2**: **Post**: "Just had a really tough day at work, but I'm trying to stay positive!" **Classification**: non self-harm **Rationale**: The post expresses frustration but lacks any direct mention of self-harm. **New Post**: {new post text} **## Output** **Classification**: [self-harm / non self-harm] **Rationale**: Describes the factors and considerations that guided the classification decision.
Fine-Tuning	The single-task fine-tuning prompt for self-harm classification and the subsequent prompt for rationale generation are similar to the one depicted in Figure **??**, with the exception that they do not include CM and SI spans, as well as emoji information sourced from the CESM-100 dataset.

Table 1: Prompts for Zero-Shot, Few-Shot, and Fine-Tuning Tasks for Self-Harm Classification and Rationale Generation

Label of synthetic post	Prompt
self-harm	**## Instruction** Create a Reddit post where the author reflects on their experiences with self-harm. The post should express a real person's emotions and thoughts, incorporating past incidents, current struggles, or stories related to their journey. Encourage the author to discuss their feelings, the impact on their life, or how they've tried to cope with their condition. The language should be varied and natural, avoiding repetition. **## Input** **Example 1**: **Post**: "I keep finding myself tracing the scars on my skin, a reminder of the war inside my head. The urge to self-harm is like a shadow that never leaves, lurking in the corners of my mind. It's a battle between wanting to feel something and wanting the pain to stop. Sometimes the blade feels like the only friend who understands. But deep down, I know it's not the answer. Trying to hold on to hope like a fragile thread in a storm. One day at a time, one breath at a time. We're warriors fighting invisible battles, and our scars tell stories of survival." **serious intent spans**: ["tracing the scars on my skin","urge to self-harm","the blade feels like the only friend"] **Rationale**: The reference of "urge to self-harm" and "the blade feels like the only friend" is unquestionably an indication of the author's genuine resolve to cause physical harm to themselves. **Example 2**: **Post**: "It's been years since I last hurt myself, but yesterday, I had the strongest urge to do it again. I found an old blade in my drawer and just sat there holding it, trying to convince myself not to use it. I didn't, but it was so hard." **serious intent spans**: ["since I last hurt myself","strongest urge to do it again"] **Rationale**: The use of phrases like "strongest urge to do it again" by the user exhibit a definite intention of causing self-injury, combined with the context of the post. **## Output**
Non-self harm	**## Instruction** Create a Reddit post containing phrases which indicate self harm in a joking or casual way. Phrases will be containing the self-harm intentions in a sarcastic way where the author is not actually going to harm himself or herself. It can show the irritation, stress, anger or disgrace just to express the situation. The author should be in a light mood to express his/her thoughts. **## Input** **Example 1**: **Post**: "Great, my boss just gave me another pointless task. Guess I'll just bang my head against the wall." **Casual intent spans**: ["bang my head against the wall"] **Rationale**: The mention of "bang my head against the wall" is a violent expression suggesting intention to hurt themselves, but,with the context of the situation, it is clearly just a metaphorical expression for frustration and does not show an actual desire for self harm. **Example 2**: **Post**: "Lost all my progress because of a glitch. Might as well just throw myself off a cliff." **Casual intent spans**: [" throw myself off a cliff."] **Rationale**: The mention of "throw myself off a cliff" shows exasperation of author about loosing his work, but it is clear looking at the emojis that it is definitely not said in a serious intent of harming themselves. **## Output**

Table 2: Prompts for generating synthetic self-harm and non-self-harm samples in our dataset