Hassles and Uplifts Detection on Social Media Narratives

Anonymous Submission

Abstract

Hassles and uplifts provide key psychological information about individuals' reactions to daily stressful situations. Identifying and collecting this information poses challenges that conventional sentiment analysis cannot fully resolve. To address this, we introduce a novel task called Hassles and Uplifts Detection (HUD) and benchmark various language models on a dataset sourced from a private social media platform. Our findings indicate that existing LLMs may not yet be fully reliable for HUD, as several key aspects require further attention. Additionally, we propose an approach to demonstrate the transferability of experimental results, overcoming the common challenge of directly publishing private datasets in the mental health domain.

1 Introduction

Mental health problem is a serious global challenge, with nearly a billion people living with a mental disorder in 2019, causing significant challenges in all aspects of life (World Health Organization, 2022). One key methodology psychologists use to study and solve mental health problem is through the analysis of hassles and uplifts in individuals' daily lives.

Hassles and uplifts are minor daily life incidents, with hassles highlighting sources of stress and uplifts offering moments of positivity that can buffer against challenges (Kanner et al., 1981; Davydov et al., 2010; Wright et al., 2020). Analyzing hassles and uplifts offers crucial insights into how individuals handle daily challenges, regulate emotions, and build resilience. This knowledge also helps psychologists to identify changes in individuals' moods (Tsakalidis et al., 2022), uncovering diverse coping strategies employed by different individuals, aiding in the prevention of mental health crises across large populations (McEwen, 2004; Bouteyre et al., 2007; Fisher, 2010; Falconier et al., 2015; Zheng et al., 2023).

Psychology and social science research have investigated social media as a data source to uncover complex emotional dynamics among people (Naslund et al., 2020; Wongkoblap et al., 2017). Several studies have shown that social media data can be leveraged to identify stress (Turcan and McKeown, 2019), detect depressive disorder (De Choudhury et al., 2013), analyze user sentiment or emotion (Zhang et al., 2024), estimate suicide risks (O'Dea et al., 2015; Chen et al., 2024), and more. However, the potential to detect and study hassles and uplifts within social media data has not yet been explored. Manual methods—such as questionnaires (Haddadi and Besharat, 2010) and diary studies (Almeida, 2005)—are impractical at large scale due to their high cost and time demands, highlighting the need for a reliable and efficient automated solution.

Unlike sentiment analysis, which primarily determines the polarity of an expressed emotion (e.g., "*I am nervous*", indicating a negative sentiment), HUD focuses on identifying specific daily life incidents that elicit positive or negative feelings. For instance, in "*Having an exam tomorrow makes me ugh #Nervous*", the phrase conveys not just a negative emotion but also the triggering event—an upcoming exam—making it a hassle. In addition, as detailed later in this paper, many social media posts may convey **mixed** hassles and uplifts, which cannot be identified through sentiment analysis. These distinctions make HUD a more nuanced challenge, requiring an understanding of both subjective emotional responses and their contextual triggers. In this paper, we introduce HUD as a novel NLP task and propose a two-step framework to address the limitations of conventional sentiment analysis. Our contributions include:

1. We propose a novel NLP task and establish a

framework for automated ***Hassles and Uplifts Detection*** (HUD).

2. We propose an approach that leverages psychometric analysis of language use in private datasets to address the common challenges of transferability and reproducibility in research related to mental health. These challenges arise because releasing sensitive mental health data alongside experimental results are often prohibited due to privacy and ethical concerns.

3. We propose an HUD data annotation guideline which covers 11 types of minor incidents individuals may encounter in daily life, offering a robust evaluation for automated HUD implementation.

4. We benchmark HUD by experimenting with various language models and configurations.

5. We conduct a qualitative analysis on benchmark predictions, highlighting that HUD offers unique advantages for assisting mental health research over conventional sentiment analysis and identifying key challenges that need to be addressed in the future.

2 Data Acquisition and Annotation

We use data provided by the owners of the Vent platform (Vent Co, 2015-2019). The original data contained 107 million posts, from which we construct and carefully annotate an HUD dataset of 500 English posts. Due to the terms of the agreement with the private data provider and the sensitive nature of the data, the HUD dataset cannot be released to the public. Instead, we provide a comprehensive description of the data acquisition and annotation guidelines. Additionally, in Section 6, we apply psychometric analysis to compare the HUD dataset with other open-source datasets, focusing on the use of cognitive process words—a key linguistic feature that highlights how individuals use language to express hassles and uplifts and manage their emotions.

2.1 Data Resource Overview

Vent is a social media platform functioning as a social diary, enabling individuals to express their feelings without restrictions. Each post includes metadata, such as a unique user identifier (user ID), a post identifier (post ID), an optional group identifier (group ID) for posts shared in specific discussion groups (e.g., "`University`"), and a binary flag to indicate explicit content.

Vent users must select a special tag to reflect their subjective feeling, such as "`sugar-rushed`", "`amused`". The tag selection is unregulated, meaning the same tag, like "`Rockin`", could either refer to a rock song or indicate that the individual is in a positive energetic state. These tags provide contextual information (Malko et al., 2021) which is treated similar to hashtags and appended to the end of each post.

2.2 Dataset Construction

The HUD dataset was created by first selecting a collection of 500 English posts, sampled from 200 unique users, as detailed below, and then having four people annotating[1] these posts. Annotators evaluated each post's content to determine whether it conveyed a *hassle*, an *uplift*, a *mix* of both, or *other* (an abstract subjective self-reflection or emotional awareness without specifying the linking incident). We manually reviewed 200 randomly selected posts and observed that Vent users do not post objective or neutral statements, thus excluding the *neutral* label. We concentrated on annotating and detecting hassles and uplifts at the level of individual posts. Thus, if the same hassle was mentioned in two consecutive posts by the same individual, it was treated as two instances of hassles. Our manual inspection of the sampled instances shows that they cover a diverse range of daily activities as illustrated in Table 1 derived from (Kanner et al., 1981). We also manually verify that no instances contain explicit content.

Data Sampling Procedure

We propose and adhere to the following data sampling pipeline and human annotation guidelines for the construction of our dataset.

1. Sample user IDs by examining their post histories and select those who have used tags from Vent's `MentalHealth` collection[2] for at least 10 times. Based on clinical psychologists' recommendations and a manual review of the posts, users who frequently tag their posts

[1] Annotators were trained through a two-round trial annotation of 100 posts (50 per round), collaboratively curating guidelines and resolving disagreements for best practices.

[2] including `Struggling`, `Persistent`, `Recovering`, `Resilience`, `Mindful`, `SetBack`, `Growing`, `Trying`, `Exhausted`, `Aware`, `Grounded`, `Helpful`, `Coping`.

Category	Itemized daily minor incidents
Close-Interpersonal Relationships	navigating family relationships, driving conversation, receiving support from, or paying obligation with family members or close friends
Study or Career	engaging with colleagues, fellows and peers, customers, teachers, employers in a study/work context; reaction to work/study load, performance, deadlines
Physical Condition	reaction to physical (dis)abilities, physical appearance, physical health, lost appetite, eating disorder
Recreation	eating out, listening music, playing sports, having or ending vacation, shopping
Pets or Animals	engaging with pets or animals (harassment of cockroaches/insects is categorized under Environment)
Environment	reaction to air quality, sound, living conditions, weather, harassment of cockroaches/flies/mice
Substance Use	taking drugs, drinking alcohol, misusing medication, smoking
Social Engagement	social media interaction, community, church, non-friend engagement
Healthcare Support	reaction to therapy or medical treatment, engaging with therapists, visiting hospitals, prescriptions
Finance	reaction to bills, salary, paying for necessities, investment, affordability
Other Activities-of-Daily Life	housework, cooking, commuting, sleep, general eating, waiting for delivery, other routine activities

Table 1: Catalog of daily minor incidents with itemized sub-incidents.

with `MentalHealth` labels are more likely to share content involving hassles and uplifts than those who rarely use these tags.

2. Sample user IDs from the previous step with *moderate* positive and negative sentiment polarity range in their post content, as these users are more likely to share varied hassles and uplifts over time. We identified these users by applying an "*off-the-shelf*" sentiment analyzer (Camacho-Collados et al., 2022), which calculates the polarity score of each post, ranging from -1 (negative) to 1 (positive). We define a person as having a moderate flow of positivity and negativity in their posting history if the $mean$ polarity score minus one standard deviation (std) for all their posts is smaller than 0 and the $mean$ polarity score plus std is greater than 0 ($mean - std < 0$ and $mean + std > 0$).

3. Sample posts indexed by the user IDs from the previous step and further sample posts with token count ranges between 8 and 120, counted under BERT tokenizer pre-trained on Twitter (Zhang et al., 2023). The rationale for this is that posts outside this token count range are likely to be tag meme challenges, diary entries, or song lyrics, according to our observations. Tag memes and song lyrics often evoke personal affects or experiences that require comprehensive inference based on the user's background knowledge. This cannot be reliably sourced from the social media content itself or the literal expression of the post. Additionally, diary entries, while rich in information, often contain deep, introspective content that is complex and multi-faceted (Aldao, 2013), requiring further studies to break down the long text content for HUD. We thus decided to exclude these posts for our current dataset and leave them for future work.

4. Sample 500 posts from the remaining data, with some of them sampled from various discussion groups based on the group IDs, including "`Friendship Match`", "`Relationships`", "`School`", "`College & University`", "`Family`", "`Weed`", "`Dogs`", "`Cats`", "`Adulting`", "`Physical Health`", "`Parenting`", and "`Drugs`". We then manually verify the sampled posts cover diverse daily incidents (see Table 1).

Human Annotation

Four annotators (demographics in Appendix A) were independently given the same instance. The annotation involved the following steps: 1) determining if an instance describes one or more incidents that have occurred, are occurring, or will occur to the post's creator. Instances that are solely self-reflective without a clear specification of any incident is labeled as a "*non-incident*"; 2) assessing if an incident(s)-containing post conveys subjective feelings and can be identified as either hassle(s), uplift(s), or a combination of both (Appendix E).

Following the human annotation, we formed four different combinations by selecting groups of three

annotators out of four and calculated the mean and standard deviation of inter-annotator agreement using Fleiss' Kappa score, resulting in $0.93_{\pm 0.02}$ for incident annotation and $0.89_{\pm 0.03}$ for subjective feeling annotation. This calculation helps account for human label variance among different subsets of annotators. To train machine learning models, we addressed label variance by consolidating labels when three annotators assigned the same label to instances with differing human annotations or by cross-annotator communication if there was a tie.

3 Framework for Identification of Hassles and Uplifts

We propose a two-step automation framework for HUD, separating objective incident detection from subjective feeling classification. This separation isolates errors in each layer, where any mistakes made in the objective incident detection do not affect the subjective feeling classification. This separation improves interpretability in evaluating the overall effectiveness of the HUD implementation.

Step (1): Incident Detection is formulated as a binary text classification, similar to event detection without trigger (Liu et al., 2019a). Input instances that describe one or more incidents are classified as "*has-incident*" and will be forwarded to the next step. Posts that do not describe any incidents are classified as "*non-incident*" and will be stopped from further HUD processing.

Step (2a): Subject Feeling Classification is formulated as a single-label, three-class classification task. It determines whether a "*has-incident*" instance forwarded from Step (1) conveys a subjective feeling, which forms a construct of either hassle, uplift, or a combination of both.

Step (2b): Sentiment Classification provides an alternative to Step (2a) and tests if subjective feeling detection can be accomplished through conventional sentiment classification. We use the labels *hassle* and *uplift* to evaluate against predictions of *negative* and *positive*, respectively. We exclude the neutral prediction as people do not post neutral and objective statements on Vent for reasons already being discussed in Section 2.1. Traditional sentiment analysis (Barbieri et al., 2020) assumes that a single sentiment dominates the text expression and thus cannot make predictions of the mixed sentiment of positive and negative.

We evaluate the HUD implementations at each step in the framework. Thus, we experiment in Step (2) by only selecting the instances that have been manually identified as "*has-incident*". In total, 278 out of 500 instances are selected. We created two sets of 3-fold cross-validation for Step 1 (based on 500 instances) and Step 2 (based on 278 instances) respectively (See Table 6, Appendix D).

3.1 Models and Experimental Setup

We experiment with both (1) a small language model, specifically a sentence transformer (Reimers and Gurevych, 2019) pre-trained using RoBERTa-Large (Liu et al., 2019b) (RoBERTa$_{hud}$), and (2) several foundational LLMs, including `Llama2-7b-chat` (Touvron et al., 2023) & `Llama3.1-8b-Instruct` (Dubey et al., 2024), `gemma2-2b-it` (Rivière et al., 2024), `gpt4-turbo` (OpenAI, 2023), and `o1-mini` (OpenAI, 2024). The two proprietary LLMs are hosted within our organization server, ensuring the privacy of processing user-sensitive data. The models' setup for the completion of each step is illustrated below:

Step (1&2a)—Incident detection and subjective feeling classification

- **Fine-tune RoBERTa$_{hud}$**: Fine-tune two separate RoBERTa-based sentence transformers on the incident and subjective feeling labels, using a contrastive learning framework with proven feasibility to be fine-tuned with few-shot examples (Tunstall et al., 2022).

- **Prompt-tune LLM$_{pt}$**: For Step 1, prompt-tune each LLM with 11 incident and 11 non-incident examples selected from the dataset for incident detection. For Step 2a, prompt-tune LLMs using 33 examples equally distributed across hassle, uplift, and mixed labels, also drawn from the dataset for subjective feeling classification. The chosen examples are manually verified to ensure they cover all categories of daily incidents illustrated in Table 1. Details on prompt formulation and modeling configurations are provided in Appendices B and C.

- **Instruction-tune `Llama3.1-8b-it`$_{ft}$**: Design task-specific instructions to combine with text instances to formulate a prompt (see prompt template in Appendix B) and to separately fine-tune two `Llama3` models on incident detection and subjective feeling classifi-

cation using 4-bit quantization and Low-Rank Adaptation configuration (Hu et al., 2022).

Step (2b)—Sentiment classification Apply an *off-the-shelf* sentiment analyzer (**RoBERTa**$_s$) (Barbieri et al., 2020) to classify each instance in the dataset as either negative or positive.

3.2 Evaluation Metrics

We use *Precision*, *Recall*, and F_1 score to evaluate each class individually in both Step (1) and Steps (2a & 2b) using a one-vs-rest approach. For example, when considering the class *hassle* as positive, any prediction of *hassle* is treated as a true positive, predictions of *uplift* or *mix* are treated as false negatives, and any *non-hassle* instances predicted as *hassle* are considered false positives. This approach ensures that the metric scores are not skewed by the majority class. We calculate the mean and standard deviation of these three metrics across 3 cross-validation folds for each step, respectively.

4 Results

The effectiveness of the models tested on the two-step HUD task is shown in Tables 2 and 3. We find that the small BERT-style model (RoBERTa$_{hud}$), fine-tuned specifically for incident detection, outperforms both the prompt-tuned and fine-tuned large language models (Table 2). `gpt4` with few-shot learning setting outperforms RoBERTa$_{hud}$ and other LLMs, including the `o1-mini` with more advanced reasoning capability, for subjective feeling classification. Comparing various LLMs, fine-tuning `Llama3` with task-specific instructions and LoRA configuration (`Llama3`$_{ft}$) yields improvements over few-shot in-context learning (`Llama3`$_{pt}$), except for the classification of mixed feeling (see Table 3). The proprietary `gpt4` and `o1-mini` outperform the open resource LLMs on both tasks.

Comparing the effectiveness of RoBERTa$_{hud}$ with RoBERT$_s$ in Table 3, we find that applying a sentiment analyzer to predict negative or positive instances as hassles or uplifts achieves a slightly better *Recall* than fine-tuning RoBERTa$_{hud}$ using hassle or uplift label ($p < 0.05$). In contrast, RoBERTa$_{hud}$ significantly outperforms the sentiment analyzer RoBERTa$_s$ in *Precision*, leading to an increased F_1 for classifying hassles and uplifts.

Comparing the performance of all experimented models on the classification of mixed hassles and uplifts in Table 3 reveals that identifying the *mix* category is more challenging than classifying instances that convey only hassles or uplifts individually. Fine-tuning a large language model improves the classification of *hassle* and *uplift* but becomes worse for the *mix* category, potentially due to insufficient training instances (under-fitting). The open resource LLMs, either prompt-tuned or fine-tuned, still fall short of the effectiveness (F_1) achieved by fine-tuning a BERT-style small model (RoBERTa$_{hud}$) or proprietary LLMs. An ideal pipeline for HUD involves using RoBERTa for incident detection, followed by `gpt4` for subjective feeling classification. Alternatively, RoBERTa can be applied to both stepped tasks, considering the computational cost of running `gpt4` (Appendix F).

5 General Interpretation of Results

Incident Detection: While event detection aims to identify triggers or explicit mentions of actions within the text (Liu et al., 2016; Ji and Grishman, 2008; Weng and Lee, 2011), incident detection in hassles and uplifts focuses on detecting occurrences that have a direct, observable relation to the individual's experience. This means that incident detection must distinguish between tangible incidents and abstract self-reflections or emotion awareness. For example, although the word "*felt*" is an event trigger in a reflection: "*I felt frustrated today*", this statement does not specify the incident that caused that frustration. In contrast, an incident-containing statement may be "*I felt frustrated today because of traffic*", which specifies the emotional response to an external occurrence, the traffic. This distinction is essential in hassles and uplifts detection, where incidents represent real-time interactions with one's environment, while events may sometimes refer to internal reflections without an actionable context. The experiment on Step (1) shows that fine-tuned `Llama3`$_{ft}$ model and RoBERTa$_{hud}$ outperforms a direct application of LLMs using in-context examples for detecting incidents (Table 2). We analyzed the error cases and found that these LLMs may have treated the incident detection as conventional event detection. For example, consider a statement like "*I had a mental breakdown yesterday*". It describes an internal experience, and the literal mention of "*yesterday*" may mislead LLMs to identify it as a specific event tied to a particular time. For our purposes,

Model	has-Incident			non-Incident		
	Precision	*Recall*	F_1	*Precision*	*Recall*	F_1
RoBERTa$_{hud}$	**0.90**$_{\pm 0.03}$	**0.92**$_{\pm 0.02}$	**0.91**$_{\pm 0.03}$	**0.90**$_{\pm 0.02}$	0.86$_{\pm 0.04}$	**0.88**$_{\pm 0.04}$
Llama2-7b-chat$_{zs}$	0.87$_{\pm 0.07}$	0.14$_{\pm 0.05}$	0.25$_{\pm 0.07}$	0.48$_{\pm 0.04}$	**0.97**$_{\pm 0.02}$	0.64$_{\pm 0.04}$
Llama2-7b-chat$_{pt}$	0.70$_{\pm 0.06}$	0.74$_{\pm 0.04}$	0.70$_{\pm 0.06}$	0.62$_{\pm 0.03}$	0.54$_{\pm 0.07}$	0.58$_{\pm 0.06}$
Llama3.1-8b-it$_{pt}$	0.62$_{\pm 0.07}$	0.87$_{\pm 0.06}$	0.72$_{\pm 0.05}$	0.68$_{\pm 0.14}$	0.32$_{\pm 0.09}$	0.43$_{\pm 0.09}$
Llama3.1-8b-it$_{ft}$	0.89$_{\pm 0.02}$	0.87$_{\pm 0.01}$	0.88$_{\pm 0.02}$	0.84$_{\pm 0.02}$	0.86$_{\pm 0.02}$	0.85$_{\pm 0.01}$
gemma2-2b-it$_{pt}$	0.58$_{\pm 0.03}$	0.95$_{\pm 0.02}$	0.72$_{\pm 0.02}$	0.67$_{\pm 0.07}$	0.95$_{\pm 0.02}$	0.72$_{\pm 0.02}$
gpt4-turbo$_{pt}$	0.89$_{\pm 0.06}$	0.81$_{\pm 0.02}$	0.85$_{\pm 0.04}$	0.79$_{\pm 0.01}$	0.88$_{\pm 0.06}$	0.83$_{\pm 0.03}$
o1-mini$_{pt}$	0.84$_{\pm 0.01}$	0.85$_{\pm 0.02}$	0.84$_{\pm 0.01}$	0.81$_{\pm 0.04}$	0.80$_{\pm 0.02}$	0.80$_{\pm 0.03}$

Table 2: Effectiveness of incident detection (Step 1).

Model	Hassle			Uplift			Mix		
	Precision	*Recall*	F_1	*Precision*	*Recall*	F_1	*Precision*	*Recall*	F_1
RoBERTa$_{hud}$	0.92$_{\pm 0.02}$	0.81$_{\pm 0.09}$	0.86$_{\pm 0.05}$	0.84$_{\pm 0.09}$	0.70$_{\pm 0.04}$	0.76$_{\pm 0.06}$	0.63$_{\pm 0.08}$	0.78$_{\pm 0.07}$	0.70$_{\pm 0.07}$
Llama2-7b-chat$_{zs}$	0.68$_{\pm 0.06}$	0.87$_{\pm 0.05}$	0.76$_{\pm 0.05}$	0.73$_{\pm 0.05}$	0.70$_{\pm 0.12}$	0.71$_{\pm 0.07}$	0.41$_{\pm 0.12}$	0.25$_{\pm 0.06}$	0.31$_{\pm 0.08}$
Llama2-7b-chat$_{pt}$	0.70$_{\pm 0.04}$	0.84$_{\pm 0.05}$	0.76$_{\pm 0.04}$	0.74$_{\pm 0.10}$	0.69$_{\pm 0.08}$	0.71$_{\pm 0.01}$	0.49$_{\pm 0.12}$	0.38$_{\pm 0.08}$	0.43$_{\pm 0.09}$
Llama3.1-8b-it$_{ft}$	0.74$_{\pm 0.01}$	0.85$_{\pm 0.04}$	0.80$_{\pm 0.02}$	0.66$_{\pm 0.10}$	0.86$_{\pm 0.11}$	0.75$_{\pm 0.10}$	0.46$_{\pm 0.25}$	0.21$_{\pm 0.08}$	0.29$_{\pm 0.12}$
Llama3.1-8b-it$_{pt}$	0.86$_{\pm 0.03}$	0.78$_{\pm 0.07}$	0.82$_{\pm 0.02}$	0.67$_{\pm 0.02}$	**0.93**$_{\pm 0.03}$	0.78$_{\pm 0.02}$	0.69$_{\pm 0.13}$	0.48$_{\pm 0.15}$	0.56$_{\pm 0.15}$
gemma2-2b-it$_{pt}$	0.77$_{\pm 0.03}$	0.77$_{\pm 0.09}$	0.77$_{\pm 0.06}$	0.64$_{\pm 0.04}$	0.92$_{\pm 0.08}$	0.76$_{\pm 0.05}$	0.51$_{\pm 0.20}$	0.26$_{\pm 0.08}$	0.34$_{\pm 0.11}$
gpt4-turbo$_{pt}$	0.91$_{\pm 0.03}$	**0.91**$_{\pm 0.07}$	**0.91**$_{\pm 0.03}$	0.87$_{\pm 0.09}$	0.79$_{\pm 0.02}$	**0.82**$_{\pm 0.05}$	**0.76**$_{\pm 0.01}$	0.82$_{\pm 0.07}$	**0.79**$_{\pm 0.04}$
o1-mini$_{pt}$	**0.95**$_{\pm 0.03}$	0.72$_{\pm 0.10}$	0.82$_{\pm 0.07}$	**0.90**$_{\pm 0.10}$	0.68$_{\pm 0.08}$	0.77$_{\pm 0.09}$	0.56$_{\pm 0.08}$	**0.88**$_{\pm 0.05}$	0.68$_{\pm 0.08}$
RoBERTa$_s$	0.74$_{\pm 0.03}$	0.86$_{\pm 0.06}$	0.79$_{\pm 0.02}$	0.66$_{\pm 0.04}$	0.83$_{\pm 0.03}$	0.73$_{\pm 0.04}$	-	-	-

Table 3: Effectiveness of subjective feeling classifications (Step 2a & 2b).

however, it is not an incident.

Subjective Feeling Classification: Our comparison between RoBERTa$_{hud}$ and RoBERTa$_s$ indicates that the HUD task is different from sentiment analysis (Table 3). While individuals may use a positive tone to express pure uplift, a general positive tone within the post does not always equate to uplift, as people may sarcastically convey hassles with positive tone: "*Had the most uplifting chat with my dad today, he called me a 'feminine' when I told him some seniors threw yogart*[3] *on me, making me cry at school. #Heartbroken*". Similarly, people may provide nuanced or cautiously expressed uplift: "*Well, I am not too sure if I'm too narcissism to talk about the feeling of honour in having this reward.*" This discrepancy leads to high *Recall* (0.83) but low *Precision* (0.66) when using a sentiment analyzer to detect uplift (and vice versa for hassle).

The sentiment analyzer (RoBERTa$_s$) cannot detect mixed hassles and uplifts. It assumes that a single sentiment polarity dominates the text expression in nature. This is problematic when trying to identify mixed hassles and uplifts, as posts in these cases do not have a singular polarity tone. Instead, they often contain both positive and negative elements that are not evenly distributed. For example, we observed posts where an active coping strategy—positive reframing—is presented. In such posts, the content might be predominantly negative with only a small positive element, as shown in "*I forgot to bring a scantron for my exam but luckily someone let me have one of theirs. I walked 20 minutes to this building and if I forgot a scantron I would've legit cried #Upset*". RoBERTa$_s$ predicted it as having the probability of 0.75 in negative but only 0.06 in positive. The complexity of the expressed feelings in posts with mixed hassles and uplifts lies in the coexistence of contrasting feelings—where the positive tone may mask underlying negative sentiments, or vice versa. This nuanced emotional dynamic cannot be captured by a simple balance of polarity using a sentiment analyzer. This also indicates that a HUD task, different from the conventional sentiment analysis, is needed for some mental health tasks, like studying emotion regulation strategies and resilience.

The prompt-tuned open resource LLMs performed poorly in detecting incidents and classifying all three subjective feelings. This is likely due to the novelty of the task, which the models have not encountered during their pre-

[3]This post is verbatim. The typo is in the original post.

training. Relying solely on few-shot examples and in-context learning does not adequately adapt them for HUD. In contrast, the two proprietary models, `gpt4` and `o1-mini`, showed stronger effectiveness in distinguishing subjective feelings. However, given their high costs and lower effectiveness than RoBERTa$_{hud}$ in detecting incidents in the earlier stage, proprietary LLMs may not be the most appropriate to use for HUD.

6 Psychometric Analysis of Cognitive Processes Words in Vent vs. Other Texts Resource

The purpose of applying psychometric analysis is two-fold. On one hand, we assess the similarity in language use within Vent compared to other open-source datasets, demonstrating that while the Vent data cannot be publicly disclosed, the HUD method remains applicable to other open-source datasets. On the other hand, we focus on analyzing the use of words related to cognitive process using LIWC (Linguistic Inquiry and Word Count) (Pennebaker et al., 2015). A prominent use of such words may indicate that the text resource contains rich information about individuals' responses to daily challenges, as well as their capacity for reflective thinking, emotional awareness and regulation. Applying HUD to such a resource can provide valuable insights and implications for resilience studies (Pennebaker and Chung, 2007). We examine the use of words related to cognitive process over $20,689$ randomly selected Vent posts, comparing with other text resources characterized by LIWC2015 (Pennebaker et al., 2015), including general Tweets, blogs, novels, expressively written diary, New York Times articles (NY Times), and natural speech. The expressive writing here is one form of diary studies (Pennebaker and Chung, 2007).

Vent posts have a relatively high-ratio use of cognitive process words (11.70), only slightly lower than expressive writing (12.52), but higher than Tweets (9.96) and all other text sources, such as NY Times (7.52) and novels (9.84) (Table 4). This supports our observation that the narratives of Vent posts are reflective and introspective, aligning closely with the nature of expressive writing. With respect to the specific category of words, Vent scores higher in the use of *insight* (2.27) and *causation* (1.52) compared to Tweets (1.92 for *insight* and 1.41 for *causation*), which suggests that users on Vent are more likely to make reasoning or reflection on their daily life experiences and feelings. Vent exhibits particularly high use of *discrepancy* (1.94) and *tentative* (2.98) words, which is even higher than in expressive writing. These categories often involve expressing uncertainty or conflicts, which could indicate a pattern of self-reflection or grappling with emotions. In contrast, tweets have notably lower scores in these two word groups (1.54 for *discrepancy* and 2.35 for *tentative*). Vent also shows a high rate of *differentiation* words (3.13), second only to natural speech (3.73), suggesting that users on Vent prefer to reflectively make distinctions between concepts or emotions, which may show various coping strategies.

7 Hassles and Uplifts Detection is not Sentiment Classification or Multi-Label Classification

While most of the negative predictions from the sentiment analyzer (RoBERTa$_s$) align with hassles, and positive predictions correspond to uplift labels, the mixed hassle and uplift annotations cannot be identified through traditional sentiment analysis. Yet, posts with mixed hassles and uplifts cannot be simply disregarded. We apply the two-step RoBERTa$_{hud}$ to first identify incidents and then classify them into hassle, uplift, or mixed feelings on a randomly sampled collection of $20,689$ posts. Out of these, $10,665$ were identified as "*has-incident*", with $3,175$ categorized as hassle, $3,797$ as uplift, and $3,693$ as mixed feelings. Thus nearly **30%** of the incident descriptions encountered convey mixed hassles and uplifts.

Additionally, posts with mixed hassles and uplifts often contain key information on how individuals regulate their emotions. Such nuanced posts are valuable for psychologists in studying resilience (Aldao, 2013). For example, in a post "*The exam is exhausting but I MADE THIS... I'VE DONE IT #Resilient*", the sentiment analyzer (RoBERTa$_s$) classifies it as positive because the positive tone dominates, failing to recognize that the exam itself was a hassle for this individual. The mixed hassles and uplift prediction can reveal an effective "reappraisal" coping strategy this individual has incorporated for building resilience. Therefore, HUD, designed to extract insights for studying mental resilience, should be treated differently from traditional sentiment classification.

Finally, we believe that formulating HUD as multi-label sentiment classification is not suitable.

Psychometrics	Vent	Tweets	Diary	Blogs	Novels	NY Times	Speech
Cognitive Process	11.70	9.96	**12.52**	11.58	9.84	7.52	12.27
insight	2.27	1.92	**2.66**	2.28	2.11	1.54	2.46
causation	1.52	1.41	**1.65**	1.46	1.03	1.42	1.45
discrepancy	**1.94**	1.54	1.74	1.56	1.48	0.89	1.45
tentative	2.98	2.35	2.89	2.82	2.27	1.74	**3.06**
certainty	1.41	1.43	1.51	**1.56**	1.45	0.76	1.38
differentiation	3.13	2.62	3.40	3.31	2.82	2.03	**3.73**
Total Instances	20,689	35,269	6,179	37,295	875	34,929	3,232

Table 4: Average LIWC score related to cognitive process among various data sources. The highest scores per row are presented in **bold**.

Given the high percentage of sentences with mixed hassles and uplifts (30% in our data), multi-label frameworks may distort evaluation results since arbitrarily assigning both hassle and uplift labels to all input instances can artificially inflate *Recall*. Moreover, in real-world applications such as mental health, psychologists require a clear distinction between mixed and singular states to accurately assess the nuanced interaction between hassles and uplifts.

8 Related Work

Researchers have utilized NLP techniques to tackle various mental health-related tasks. Some researchers have designed NLP approaches to automatically classify sentiment polarity or emotional states (Zhang et al., 2024; Barbieri et al., 2020), detect stress (Xu et al., 2024; Turcan and McKeown, 2019), identify ironic (Van Hee et al., 2018) or abusive (Nobata et al., 2016) expression, or detect depressive disorder (Wolohan et al., 2018) from an individual's text expression. However, no approaches have been explicitly designed to detect daily hassles and uplifts. Rather than simply categorizing a text as either positive or negative, the core information need of HUD lies in analyzing both the objective incident described and the subjective feeling it caused. This approach uncovers emotional complexity and the nuanced context of an individual's personal life, providing key insights into how individuals cope and adapt to life's challenges.

LLMs have shown effectiveness to handle a wide range of text processing tasks, but adapting them for HUD and ensuring their reliability remains unexplored. While BERT-style small language models (SLMs) demonstrate better performance over text classification than large generative language models (Bucher and Martini, 2024), the effectiveness of SLMs requires task-specific fine-tuning, such as those used in sentiment analysis (Barbieri et al., 2020). The reliance on large annotated training sets presents challenges for real-world applications, as consistently annotating task-specific datasets for every unique context is impractical. For example, the instances used to adapt SLMs in workplace satisfaction studies (Fisher, 2010) can differ significantly from those in studies on university student stress (Bouteyre et al., 2007). Therefore, studying feasible approaches to adapt SLMs with a small training set for HUD is needed.

9 Conclusion

We introduced a novel task to automate the detection of hassles and uplifts within social media narratives, presenting its unique advantage that conventional sentiment analysis cannot fully resolve. Through a series of experiments using various language models, we benchmarked the effectiveness of HUD and found that these models, particularly large language models (LLMs), are not yet fully reliable for this task. Key challenges persist, especially in distinguishing tangible life incidents from more abstract forms of self-reflection or emotional awareness. Furthermore, we proposed an approach that utilizes psychometric analysis to compare language use between public and private datasets, offering a way to overcome the common challenge of releasing sensitive mental health data alongside with experimental results.

In the future, we aim to extend HUD to other, especially public, datasets. We also plan to verify the effectiveness of HUD on downstream tasks, such as for assisting the identification of emotion regulation strategies or "*moments of change*" in individuals' moods over time (Tsakalidis et al., 2022).

Limitations

In this paper, we simplify the detection of hassles and uplifts within the scope of a single post. However, we observed cases where an individual may express follow-up subjective feelings towards an incident mentioned in earlier posts. Since the latter posts only convey subjective feelings without specifying the aforementioned incident, our framework will not treat such posts as containing information of hassles but rather a post of pure emotional awareness.

We temporarily excluded the detection of hassles and uplifts on tag memes, journal entries, and song lyrics, because detection on such data sources requires inference on individuals' background knowledge, which cannot be reliably sourced from social media content itself or the literal expression of the posts.

We also excluded for now the detection of neutral feelings. From our observations, people generally did not post on Vent about incidents that evoked only neutral emotions. However, we have noted instances where posts initially mention a hassle but shift toward a neutral or moderate tone after self-coping, such as, "*Even if I finish my drawings and paintings in time, I have absolutely no idea how to get to this university. I hope this will be a nice week. #Anxious*" Having said that, the primary goal of HUD is not to identify reflective outcomes but to focus on the direct associations between incidents and the subjective feelings they evoke as hassles or uplifts.

Our data is in English, and our results are limited to one platform. The data is also private due to its sensitivity (mental health) and potential risk of having identifiable information.

Ethical Concerns

We have obtained ethics approval from our respective institutions to use the data provided through the Vent platform for research purposes within restricted terms: the data is not to be shared beyond our research team and it must be stored in a secure setup within the organization. We have used examples of the posts that are not identifiable. The annotators had no access to the original posts or the authors' identities.

Potential Risks

The tested language models carry the risk of producing biased and potentially harmful predictions. They do not satisfy clinical standards to deliver accurate assessments, and their inaccurate or insensitive responses could downplay individuals' struggles or even exacerbate emotional distress. To safeguard the privacy and consent of data providers, information about the cultural and demographic backgrounds of the users who generate the data was not collected. However, this lack of context can result in misunderstandings of culturally specific emotional expressions, leading to alienating or inappropriate outcomes. We note, however, that our aim is not to provide automated mental health apps but to support psychologists in their work.

References

Amelia Aldao. 2013. The future of emotion regulation research: Capturing context. *Perspectives on Psychological Science*, 8(2):155–172.

David M Almeida. 2005. Resilience and vulnerability to daily stressors assessed via diary methods. *Current Directions in Psychological Science*, 14(2):64–68.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics.

Evelyne Bouteyre, Marion Maurel, and Jean-Luc Bernaud. 2007. Daily hassles and depressive symptoms among first year psychology students in france: The role of coping and social support. *Stress and Health: Journal of the International Society for the Investigation of Stress*, 23(2):93–99.

Martin Juan José Bucher and Marco Martini. 2024. Fine-tuned 'small' LLMs (Still) Significantly Outperform Zero-shot Generative AI Models in Text Classification. *arXiv preprint arXiv:2406.08660*.

Jose Camacho-Collados, Kiamehr Rezaee, Talayeh Riahi, Asahi Ushio, Daniel Loureiro, Dimosthenis Antypas, Joanne Boisson, Luis Espinosa Anke, Fangyu Liu, and Eugenio Martínez-Cámara. 2022. TweetNLP: Cutting-Edge Natural Language Processing for Social Media. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–49.

Jiyu Chen, Vincent Nguyen, Xiang Dai, Diego Molla, Cecile Paris, and Sarvnaz Karimi. 2024. Exploring instructive prompts for large language models in the extraction of evidence for supporting assigned suicidal risk levels. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 197–202.

Dmitry M Davydov, Robert Stewart, Karen Ritchie, and Isabelle Chaudieu. 2010. Resilience and mental health. *Clinical Psychology Review*, 30(5):479–495.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 7, pages 128–137.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Mariana K Falconier, Fridtjof Nussbeck, Guy Bodenmann, Hulka Schneider, and Thomas Bradbury. 2015. Stress from daily hassles in couples: Its effects on intradyadic stress, relationship satisfaction, and physical and psychological well-being. *Journal of Marital and Family Therapy*, 41(2):221–235.

Cynthia D Fisher. 2010. Happiness at work. *International Journal of Management Reviews*, 12(4):384–412.

Parvaneh Haddadi and Mohammad Ali Besharat. 2010. Resilience, vulnerability and mental health. *Procedia-Social and Behavioral Sciences*, 5:639–642.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Heng Ji and Ralph Grishman. 2008. Refining event extraction through cross-document inference. In *Proceedings of ACL-08: HLT*, pages 254–262.

Allen D Kanner, James C Coyne, Catherine Schaefer, and Richard S Lazarus. 1981. Comparison of two modes of stress measurement: Daily hassles and uplifts versus major life events. *Journal of Behavioral Medicine*, 4:1–39.

Shulin Liu, Yang Li, Feng Zhang, Tao Yang, and Xinpeng Zhou. 2019a. Event detection without triggers. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 735–744.

Shulin Liu, Kang Liu, Shizhu He, and Jun Zhao. 2016. A probabilistic soft logic based approach to exploiting latent and global information in event classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 1, pages 2993 – 2999.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Anton Malko, Cecile Paris, Andreas Duenser, Maria Kangas, Diego Molla, Ross Sparks, and Stephen Wan. 2021. Demonstrating the reliability of self-annotated emotion data. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 45–54.

Bruce S McEwen. 2004. Protection and damage from acute and chronic stress: allostasis and allostatic overload and relevance to the pathophysiology of psychiatric disorders. *Annals of the New York Academy of Sciences*, 1032(1):1–7.

John A Naslund, Ameya Bondre, John Torous, and Kelly A Aschbrenner. 2020. Social media and mental health: benefits, risks, and opportunities for research and practice. *Journal of Technology in Behavioral Science*, 5:245–257.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*, pages 145–153.

Bridianne O'Dea, Stephen Wan, Philip J Batterham, Alison L Calear, Cecile Paris, and Helen Christensen. 2015. Detecting suicidality on Twitter. *Internet Interventions*, 2(2):183–188.

OpenAI. 2023. GPT4. Version: firstcontact-gpt4-turbo 2023-03-15-preview.

OpenAI. 2024. o1-mini. Version: o1-mini 2024-02-15-preview.

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of LIWC2015.

James W Pennebaker and Cindy K Chung. 2007. Expressive writing, emotional upheavals, and health. *Foundations of Health Psychology*, pages 263–284.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Morgane Rivière, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, et al. 2024. Gemma 2: Improving Open Language Models at a Practical Size. *CoRR*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Adam Tsakalidis, Federico Nanni, Anthony Hills, Jenny Chim, Jiayu Song, and Maria Liakata. 2022. Identifying moments of change from longitudinal user text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4647–4660.

Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. 2022. Efficient few-shot learning without prompts. *arXiv preprint arXiv:2209.11055*.

Elsbeth Turcan and Kathleen McKeown. 2019. Dreaddit: A Reddit Dataset for Stress Analysis in Social Media. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pages 97–107.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018. SemEval-2018 task 3: Irony detection in English Tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 39–50, New Orleans, Louisiana. Association for Computational Linguistics.

Vent Co. 2015-2019. Vent official website. https://www.vent.co/.

Jianshu Weng and Bu-Sung Lee. 2011. Event detection in twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 5, pages 401–408.

JT Wolohan, Misato Hiraga, Atreyee Mukherjee, Zeeshan Ali Sayyed, and Matthew Millard. 2018. Detecting linguistic traces of depression in topic-restricted text: Attending to self-stigmatized depression with NLP. In *Proceedings of the First International Workshop on Language Cognition and Computational Models*, pages 11–21.

Akkapon Wongkoblap, Miguel A Vadillo, and Vasa Curcin. 2017. Researching mental health disorders in the era of social media: systematic review. *Journal of Medical Internet Research*, 19(6):e228.

World Health Organization. 2022. World mental health report: transforming mental health for all: executive summary. In *World mental health report: transforming mental health for all: executive summary*. World Health Organization.

Aidan GC Wright, Elizabeth N Aslinger, Blessy Bellamy, Elizabeth A Edershile, and William C Woods. 2020. Daily Stress and Hassles. *The Oxford Handbook of Stress and Mental Health*, page 27.

Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K Dey, and Dakuo Wang. 2024. Mental-LLM: Leveraging large language models for mental health prediction via online text data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(1):1–32.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Pan, and Lidong Bing. 2024. Sentiment analysis in the era of large language models: A reality check. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3881–3906.

Xinyang Zhang, Yury Malkov, Omar Florez, Serim Park, Brian McWilliams, Jiawei Han, and Ahmed El-Kishky. 2023. Twhin-bert: A socially-enriched pre-trained language model for multilingual tweet representations at Twitter. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5597–5607.

Hao Zheng, Eric M Cooke, Kehan Li, and Yao Zheng. 2023. Capturing hassles and uplifts in adolescents' daily lives: Links with physical and mental well-being. *Journal of Youth and Adolescence*, 52(1):177–194.

A Annotator Demographics

The four annotators of the HUD dataset are highly educated and work in English-speaking countries. They come from different gender and linguistic backgrounds.

B Prompt Template for LLMs

The exact prompts used in our experiments for LLMs are shown below. The prompts are formulated with chat template with respect to each LLM.

- Prompt Template for **Incident Detection**:
  `You are a binary text classifier. Does the following text describes an incident or not? A self-reflective process is not incident. Requirement: Only answer 1 as incident and 0 as non-incident. For example, {{few-shot examples}} {{TEXT content to be processed}}`

- Prompt Template for **Subjective Feeling Classification**:
  `You are a psychologist and a text classifier. Does the following text describes uplift (1), hassle (-1), or the mix of both (0)? Requirement: Only answer -1 as hassle, 1 as uplift, and 0 as mixture of both. For example, {{few-shot examples}} {{TEXT content to be processed}}`

C Machine Learning Configuration

We list the information on the configurations of open resource LLMs in Table 5.

params	value
GPU	NVIDIA RTX 3500 Ada
context size	512
temperature	0.01
quantization	4-bits
LoRA rank	32
LoRA alpha	64

Table 5: The environment setting and parameters for QLoRA fine-tuning and prompt-tuning `Llama2&3` and `Gemma2`.

D Statistics of Cross Validation Dataset

The statistics of three-fold cross validation set is shown in Table 6.

Fold	Incident Detection		Subj Feeling Detection		
	Incident	Non-incident	hassle	mix	uplift
$cv1_{eval}$	93	73	37	28	28
$cv2_{eval}$	93	76	38	27	28
$cv3_{eval}$	92	73	35	27	30
total	278	222	110	82	86

Table 6: The count of instances used for evaluation per the cross-validation fold.

E HUD Annotation Guideline

E.1 Disclaimer of Risks

As an annotator working with social media data involving daily hassles and uplifts, you may encounter content that could be emotionally sensitive, distressing, or explicit. The data you will annotate may include expressions of frustration, anger, sadness, or other emotional states, as well as positive or uplifting content. While efforts have been made to filter harmful material, some posts may still include pornographic, explicit, or otherwise offensive content, which could be unsettling or distressing depending on your personal sensitivities and experiences. If you encounter content that you find distressing, we encourage you to notify the project team.

Participation in this annotation project is voluntary, and you have the right to withdraw at any time. By proceeding, you acknowledge the potential emotional and psychological risks involved, including exposure to explicit material, and confirm that you are aware of available resources to manage your well-being during this task.

E.2 Task description

Annotating social media posts that either convey a hassle, uplift, or a mixed of both by the content creator.

E.3 Instructions

1. Read the post content and the self-reported hashtag.

2. Decide whether the post conveys daily minor incident(s) of either hassles, uplifts, or mix of both.

3. Select both *hassles* and *uplifts* label if you think a post describes incident(s) and convey a mixed of both.

4. Leave the cell empty if the text does not describe any incidents.

5. Select *unknown* if you cannot decide the subjective feeling conveyed by the post content (rare case).

6. Leave necessary comment in the cell indexed by the Comment column.

E.4 Instruction on Distinguishing Incident and Non-incident Instance

There is no universal definition of what describes an incident or non-incident. In this task, we define an incident to be a specific occurrence involving participants. An incident is something that happened in the past, is happening now, or is expected to happen in the future. An incident can frequently be described as a change of state. In contrast, non-incident instances are likely to be solely containing self-reflection or emotional awareness. Examples of two non-incident instances:

> *I broke down crying, i am really sad. i never thought i could feel this much again, but it seems like i was wrong. i feel everything and it was too much, it feels like my heart is breaking all over again. this time im truly alone again. feels like 2017 all over again :') #Sad*

> *Anxious as hell today. Ugh, hate that feeling. But, I won't let it control me. It's not gonna stop me from doing all the things I want to do. Ever. #Struggling*

E.5 Instruction on Distinguishing Major and Minor Incident

There is no clear boundary between major or minor incidents, as it depends on the subjective scope of an individual. Major incidents are less frequent and cause long-term impact to the individual, such as being diagnosed with cancer or job loss. For simplification, we annotate both posts as has-incident regardless of major or minor incidents it conveys.

E.6 Definition of Hassles and Uplifts

Both hassles and uplifts are daily minor incidents. **Hassles** conveys experiences and conditions of daily living that have been appraised as negative and harmful or threatening to the post creator's well-being. **Uplifts** conveys experiences and conditions of daily living that have been appraised as positive or favorable to the post creator's well-being. Below are examples of a hassle, an uplift, and a mixed instances:

> **Hassle instance**: *Having an exam tomorrow makes me nervous ugh #Nervous*

> **Uplift instance**: *So, I'm a mother again.....to a new kitten. #Optimistic*

> **Mixed instance**: *Thank God. Tomorrow I don't need to be at work until 1pm. Which means I can sleep in. I really fucking need it after this week. #Coping*

F Estimation of Computational Cost

The approximate GPU hours (NVIDIA RTX 3500 Ada) for a few-shot application of open-resource LLMs or supervised LoRA fine-tuning `Llama3` with 4-bits quantization are all within 0.5 hours. The approximation of cost for running proprietary LLMs is shown in Table 7.

Model	Cost
`gpt4-turbo`	$\approx$ 18.40
`o1-mini`	$\approx$ 2.02

Table 7: The estimation of cost (USD) for running proprietary LLMs on the HUD dataset. The estimation is based on the count of input and output tokens.