Hassles and Uplifts Detection on Social Media Narratives

Jiyu Chen[1], **Sarvnaz Karimi**[1], **Diego Mollá**[2,1],
Andreas Duenser[1], **Maria Kangas**[2], **Cécile Paris**[1,2]

[1]CSIRO Data61, Australia
firstname.lastname@csiro.au

[2]Macquarie University, Australia
diego.molla-aliod@mq.edu.au

Abstract

Hassles and uplifts are psychological constructs of individuals' positive or negative responses to daily minor incidents, with cumulative impacts on mental health. These concepts are largely overlooked in NLP, where existing tasks and models focus on identifying general sentiment expressed in text. These, however, cannot satisfy targeted information needs in psychological inquiry. To address this, we introduce Hassles and Uplifts Detection (HUD), a novel NLP application to identify these constructs in social media language. We evaluate various language models and task adaptation approaches on a probing dataset collected from a private, real-time emotional venting platform. Some of our models achieve F_1 scores close to 80%. We also identify open opportunities to improve affective language understanding in support of studies in psychology.

1 Introduction

Hassles and uplifts are psychological constructs denoting daily minor events that can trigger positive or negative sentiment in individuals, as reflected through their language use (Kanner et al., 1981; Wright et al., 2020). Research highlights that this information provides critical insights to the psychological studies of emotion regulation, coping, and resilience (Davydov et al., 2010; Crane et al., 2019; Falon et al., 2021; Bolger et al., 2003). Hassles and uplifts often mark the onset of the emotion regulation process (Gross, 1998, 2015), and have been employed as anchors in psychological studies using experience sampling (Myin-Germeys et al., 2009; Stone et al., 2023). Sentiment analysis does not specifically distinguish or model incident-triggered emotions from general sentiment expressions. Table 1 presents illustrative examples highlighting several major distinctions, where detecting incident-triggered sentiment is more challenging, requires finer-grained contextual understanding, and remains underexplored: (a) expresses a negative sentiment (nervousness), but no specification of the triggering incident; (b) is a negative response (nervousness) triggered by a specific incident (an exam taking place in the following week); (c) expresses a negative reflection to a general aspect of life (exams); finally, (d) expresses both a negative feeling and an incident about to occur which triggers a positive response. Most sentiment analyzers would classify (a), (b) and (c) as negative; yet, only (b) expresses a hassle, by stating the incident that triggers the negativity. Example (d) expresses both a reflection of negative feeling and positive sentiment triggered by an upcoming event (an uplift). A common sentiment analyzer is likely to classify (d) as neutral, while an aspect-based one might identify both negative and positive sentiments (with different aspects), but will still miss the uplift. Specifically identifying hassles and uplifts, as opposed to only positive and negative sentiments expressed broadly in a sentence, would be useful in psychology to study resilience and emotion regulation. We thus propose the hassles and uplifts detection (HUD) task to identify such incident grounded sentiment from social media text. The HUD task can be seen as a specific type of aspect-based sentiment analysis, ABSA (Nazir et al., 2020), targeting specifically incidents that induce positive, negative, or mixes of both senti-

	Example
(a)	*I feel nervous.*
(b)	*I am nervous about the exam next week.*
(c)	*Exams always make me nervous.*
(d)	*I'm feeling really down these days, but at least good to hear my mom is visiting me next month.*

Table 1: Sentences expressing sentiment under different conditions. (a) a general negative sentiment; (b) a negative sentiment triggered by a specific incident (*hassle*); (c) a negative sentiment towards a general aspect of life; (d) start with a negative sentiment, and followed by an incident triggering positive sentiment (*uplift*).

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 475–489

ments. Incidents are discrete, tangible, and temporally bounded events that trigger momentary experiences (Stone et al., 2023), distinguishing them from background circumstances or self-reflections, or from aspects identified in ABSA in domains like product reviews (Hu and Liu, 2004). The incidents captured by HUD are also more specific than events considered in general event detection in NLP (Liu et al., 2019a; Araki and Mitamura, 2018). Event detection systems would detect events in Examples (a) & (c) in Table 1, with "feel" or "make" as the event trigger. Yet these sentences express reflections, not specific incidents (past or future) triggering a negative or positive response, which is the information of interest to psychologists. HUD is also novel in disentangling hassles or uplifts from sentences with mixed sentiments, as in (d), often frequent in language, yet overlooked by existing NLP systems and required for some psychological studies.

In our study, we implement HUD on social media text as this source of data has proven value in assisting mental health research (Naslund et al., 2020; Wongkoblap et al., 2017). Prior works used such resource for stress detection (Turcan and McKeown, 2019), depression prediction (De Choudhury et al., 2013), sentiment analysis (Zhang et al., 2024), and suicide risk estimation (O'Dea et al., 2015; Chen et al., 2024). However, the detection of hassles and uplifts from social media text remains unexplored despite its practical need in psychology. We address this information need by exploring the effectiveness of existing language modeling approaches for HUD in a data-scarce setting. We also identify several remaining challenges. Our contributions are:

1. Introducing Hassles and Uplifts Detection (HUD) as a novel NLP task that identifies affective responses to specific events (incidents), grounded in psychological theory.

2. Developing detailed annotation guidelines covering a diverse set of categories of daily incidents to ensure diversity in annotation. (11 categories are included.)

3. Evaluating five state-of-the-art language models under realistic low-resource adaptation settings and revealing key limitations of existing approaches, highlighting HUD as a challenging task.

4. Performing a psycholinguistic analysis of a HUD dataset to characterize the language features associated with introspective sharing of hassles and uplifts. Although the dataset cannot be released due to privacy constraints, we compare its stylistic properties with multiple public and private corpora, enabling future researchers to construct and validate comparable datasets under similar ethical limitations.

2 Data Acquisition and Annotation

We obtained data through a formal agreement with the developer of the Vent platform (Vent Co, 2015-2019). The entire dataset contains 107 million posts, from which we randomly curated and annotated a subset of 650 English-language posts to construct a probing dataset. Due to the sensitivity of the content and restrictions imposed by our data-sharing agreement, the HUD dataset cannot be made publicly available. To support transparency and facilitate future research, we (1) publicly share the data acquisition protocol and annotation guidelines (Section 2.2); and (2) conduct a psycholinguistic analysis using the LIWC tool (Pennebaker et al., 2015) to characterize the linguistic patterns of the dataset and compare them with multiple publicly available social media corpora (Section 6).

2.1 Data Source Overview

Vent is a social media platform featuring inward-focused, self-reflective content that closely resembles ecological diaries of emotional experiences, and has proven value for mental health research (Turcan et al., 2021; Malko et al., 2023). 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.

Each post includes a single self-selected emotion tag, such as "`overwhelmed`" or "`amused`". Prior work has shown that these tags serve as valuable indicators of users' emotional valence (Malko et al., 2021), especially given the inherent difficulty of inferring someone's sentiment and emotion solely from a third-party observer perspective. However, because the tagging process is not standardized, tags can be ambiguous; for instance, "`Rockin`" may refer either to a musical reference or an energized state. To preserve this context and minimize ambi-

guity, we concatenate each tag to the corresponding post text during pre-processing.

2.2 Dataset Construction

The HUD dataset was created by first randomly sampling 650 English posts as detailed below, and then having four people annotate these posts. Annotators were trained through a two-round trial annotation of 100 posts (50 per round), collaboratively curating guidelines and resolving disagreements for best practices. Annotators evaluated each post's content to determine whether it conveyed a *hassle*, i.e., a negative reaction to an incident, an *uplift*, i.e., a positive reaction to an incident, a *mix* of both, or *other* (general emotional expression or reflection on life situation without specifying any incident). We reviewed the sampled posts and observed that Vent users do not post objective or neutral statements as they use a self-selected emotional tag to indicate non-neutrality, thus excluding the *neutral* label. Our visual inspection of the sampled posts shows that they cover a diverse range of daily activities (see Table 6, Appendix A), derived from (Kanner et al., 1981).

Data Sampling Procedure

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

Step 1 Sample user IDs by examining their post histories and select those who have used tags from Vent's `MentalHealth` collection[1] at least 10 times. Based on clinical psychologists' recommendations and visual inspection of the posts, users who frequently use `MentalHealth` labels are more likely to share content involving hassles and uplifts than those who rarely use these tags.

Step 2 Randomly sample 3500 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 sentiment in their posting behavior if the mean polarity score (μ) and standard deviation (σ) on all of their posts satisfy $\mu - \sigma < 0$ and $\mu + \sigma > 0$, respectively.

Step 3 Sample 650 posts from the previous step, 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 remove tag memes, posts where the length exceeded 250 words, and two instances of song lyrics (inferred by the share links to the song in the post) and randomly resample replacement posts. Tag memes and song lyrics often evoke personal affects or experiences that require comprehensive inference based on the user's background knowledge. These cannot be reliably sourced from the social media content itself or the literal expression of the post. Additionally, long posts, while rich in information, often contain deep, retrospective content that is complex and multi-faceted (Aldao, 2013), requiring further studies to break down the long text content for HUD. Therefore, as advised by one of the authors of this paper, a clinical psychologist, we exclude these posts for our current task setting and leave them for future work.

Human Annotation

The team consulted with two psychologists (Dr Kangas and Dr Duenser, co-authors of this paper) for developing the guidelines. Four annotators, educated and working in English-speaking countries, independently annotated all the posts. All annotators underwent training and two rounds of trial annotations prior to formal labeling. The annotation (full guideline in Appendix F) followed two steps: (1) Determining if an instance describes one or more incidents that have occurred, are occurring, or will shortly occur in the life of the post's creator. Otherwise, annotate as "*non-incident*": in such cases, the post only provides a generic sentiment expression or self-reflection, possibly on chronic experiences. (2) Assessing if the incident described in the post triggers either positive, negative, or a mixture of both sentiments for the individual. For each instance, we iteratively paired four groups of labels from three out of four annotators and calculate the Fleiss' Kappa score. This iterative pairing enables the calculation of mean and standard deviations of the agreement score. Overall, we measured

[1]This collection of tags includes `Struggling`, `Persistent`, `Recovering`, `Resilience`, `Mindful`, `SetBack`, `Growing`, `Trying`, `Exhausted`, `Aware`, `Grounded`, `Helpful`, and `Coping`.

$0.93_{\pm 0.02}$ for incident annotation and $0.89_{\pm 0.03}$ for triggered sentiment. The score proves that, despite HUD being a subjective language understanding task, human annotators can still reliably distinguish these constructs from general emotional expressive content, as directed by our annotation guidelines. After having checked the inter-annotator agreement, we assigned a ground-truth label for each instance by selecting the one assigned by at least three annotators or by cross-annotator communication if there was a tie.

The final dataset had 323 out of 650 (49.7%) posts annotated as *has-incident*. From those, 128 (39.6%) were hassles, 112 (34.7%) uplifts and 83 (25.7%) were mixed.

3 A Framework for Identification of Hassles and Uplifts

We propose a two-step pipeline for the HUD task: (1) incident detection, and (2) the classification of the sentiment triggered by those incidents.

Step 1: Incident Detection is modeled as a binary text classification. Input posts that describe one or more incidents are classified as "*has-incident*" and passed on to the next step to classify the triggered sentiment(s). Posts that do not specify any incident are labeled as "*non-incident*" and excluded from further processing.

Step 2: Incident-triggered Sentiment Classification is modeled as a single-label, three-class classification. For each post predicted as "*has-incident*", this step classifies the emotional valence elicited by the described incident as either positive (uplift), negative (hassle), or mixed.

Baseline: We establish an integrated baseline by contrastively fine-tuning (Tunstall et al., 2022) a RoBERTa-Large-based (Liu et al., 2019b) sentence transformer (Reimers and Gurevych, 2019) (denoted as RoBERTa$_i$) as a single-label, four-class classifier over the HUD labels: non-incident, hassle, uplift, and mix. We compare this baseline with the end-to-end two-step pipeline we propose (denoted as RoBERTa$_e$) to assess cascading errors. The pipeline is constructed by fine-tuning two RoBERTa models, one for each step, which correspond to the two RoBERTa$_{ft}$s described in Section 3.1.

We also perform an isolated evaluation for each pipeline component. Step 1 is evaluated on all posts, whereas Step 2 is evaluated only on posts labeled as "*has-incident*". We perform two 5-fold cross-validation on each step, respectively. (See the statistics of each fold in Table 8, Appendix D.)

3.1 Models and Experimental Setup

We evaluate fine-tuned sentence transformers (RoBERTa$_{ft}$), and several LLMs for each pipeline step, including `Llama3.1-8b-Instruct` (Dubey et al., 2024), `gemma2-2b-it` (Rivière et al., 2024), `gpt4-turbo` (OpenAI, 2023), and `o1-mini` (OpenAI, 2024). This selection balances between small and large language models, as well as open-source and proprietary models, to ensure the robustness of our evaluation. HUD shares two common challenges in the application of NLP methods to mental health research: (1) the limited availability of ground-truth labeled data; and (2) the privacy and time constraints that hinder large-scale data acquisition. Therefore, we adopt task adaptation strategies that are proven effective in low-resource settings:

Contrastive fine-tune RoBERTa$_{ft}$: Fine-tune two separate RoBERTa-based pre-trained sentence transformers on the ground-truth incident and triggered sentiment label respectively, using a contrastive learning framework (Tunstall et al., 2022).

In-context prompting LLM$_{pt}$: Prompt LLMs using all posts and corresponding ground-truth labels from each training fold as in-context learning examples, and instruct LLMs to make predictions on the posts in the validation fold. To assess the impact of prompt variation and the few-shot sample selection, we initially tested multiple prompt designs with random selection of few-shot samples without replacement on a subset of 100 posts and observed negligible changes in the results. Details on prompt engineering and modeling configurations are provided in Appendices B and C.

Supervised fine-tune `Llama3.1-8b-it`$_{ft}$: Combine task-specific instructions with post content to formulate an instructive prompt (see prompt template in Appendix B), and separately fine-tune `Llama3` models on each cross-validation fold and for each HUD step using 4-bit quantization and Low-Rank Adaptation (Hu et al., 2022). While we acknowledge the existence of additional low-resource task adaptation methods, the chosen strategies are representative and sufficient to justify the feasibility and current limitations of NLP approaches in supporting HUD.

3.2 Evaluation Metrics

We use *Precision*, *Recall*, and F_1 scores to evaluate each class individually in both HUD steps using a one-vs-rest approach, ensuring the metric scores are not skewed by the majority class. 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* posts predicted as *hassle* are considered false positives.

Once we obtain the model predictions for each fold in our cross-validation setup, we apply a non-parametric bootstrap procedure (Efron, 1992) to determine the confidence intervals. Specifically, we perform 100 bootstrap iterations on the set of prediction–label pairs from the entire dataset. In each iteration, we sample with replacement from the original prediction set to create a bootstrap sample of the same size. The *Precision*, *Recall*, and F_1 score, are then computed on each resampled set. We report the mean and the 95% confidence interval to represent the variability of performance estimates. This approach ensures that, despite the small size of our dataset, model performance is not unduly influenced by coincident sampling of particularly hard instances or overinflated due to sampling easier cases in the cross-validation folds. It also ensures the effectiveness translates to the entire dataset within the confidence intervals (Dror et al., 2018).

4 Experimental Results

In the incident detection step (Table 2), RoBERTa$_{ft}$ outperforms both open-resource and proprietary LLMs in F_1 score. Among the four LLMs, the two proprietary models outperform both `gemma2` and `llama3`. Regarding model size, the smaller 2B-parameter `gemma2` performs comparably to the larger 8B `llama3` on "*has-incident*" classification, but significantly underperforms on "*non-incident*" posts. Notably, supervised fine-tuning of `llama3` using LoRA does not yield notable gains over prompt-tuning and even degrades in the "*has-incident*" case. The gain of supervised fine-tuning in the "*non-incident*" cases is achieved through over-sensitivity, as reflected by overrated *Recall* and a significant drop in *Precision*. While scaling up training data might improve effectiveness, such expansion is impractical in mental health contexts due to resource and sensitivity constraints. Given these limitations, contrastive fine-tuning of sentence transformers appears to be the most viable adaptation strategy for incident detection.

In the sentiment classification step (Table 3), all models perform better at detecting purely hassles or uplifts than identifying mixed cases. The two prompt-tuned proprietary LLMs achieve the highest F_1. Although `RoBERTa`$_{ft}$ does not achieve the top mean score, it performs competitively with both proprietary LLMs. In contrast, `gemma2` suffers a substantial performance drop when classifying "*mix*" cases. Supervised fine-tuning of `llama3` using LoRA also results in a marked decline in F_1, indicating that even lightweight fine-tuning techniques may be ineffective to enable nuanced understanding for open-resource LLM under extremely low-resource conditions.

RoBERTa$_e$ (end-to-end) shows no statistically significant degradation compared to RoBERTa$_{ft}$ (Step 2 only), indicating minimal impact from cascading errors (Table 3). The two-step setup also outperforms the integrated model (RoBERTa$_e$ vs. RoBERTa$_i$); while RoBERTa$_i$ accurately detects incidents, it fails to distinguish subtypes (i.e., hassle, uplift, mix), supporting the effectiveness of the two-step formulation.

5 Qualitative Error Analysis

To this end, we exclude `llama3`$_{ft}$ from the qualitative error analysis. Empirical evidence in Section 4 already indicated that supervised fine-tuning of large language models yields no substantial performance gains over in-context prompting in this setting. As such, the observed errors are likely attributable not to the subtlety of affective language per se, but rather to factors related to model configuration and limitations in low-resource adaptation. Our analysis focuses on uncovering the core challenge in understanding the nuances in language expressions that involve hassles and uplifts.

Incident Detection: We categorize the 650 probing posts into three levels of difficulty based on the degree of agreement between model predictions and the ground truth. We consider easy cases to be those where either all five models are correct or only one model makes an error. Moderate cases involve two or three models that made incorrect predictions. Highly challenging cases are those where four or all five models made errors. Based on this criterion, we identify 376 easy, 219 moderate, and 55 highly challenging posts. The number of errors made on the moderate and highly challenging posts

Model	has-Incident			non-Incident		
	Precision	*Recall*	F_1	*Precision*	*Recall*	F_1
RoBERTa$_{ft}$	**0.86**$_{[.82,.90]}$	0.88$_{[.85,.91]}$	0.87$_{[.84,.89]}$	0.88$_{[.84,.91]}$	**0.86**$_{[.81,.90]}$	0.87$_{[.83,.89]}$
llama3$_{pt}$	0.57$_{[.53,.61]}$	0.88$_{[.85,.92]}$	0.69$_{[.66,.72]}$	0.75$_{[.69,.83]}$	0.34$_{[.30,.39]}$	0.47$_{[.42,.52]}$
llama3$_{ft}$	0.71$_{[.63,.78]}$	0.39$_{[.34,.44]}$	0.51$_{[.45,.55]}$	0.59$_{[.54,.63]}$	0.84$_{[.80,.89]}$	0.69$_{[.66,.73]}$
gemma2$_{pt}$	0.52$_{[.48,.56]}$	**0.95**$_{[.92,.97]}$	0.67$_{[.64,.70]}$	0.73$_{[.63,.84]}$	0.14$_{[.11,.18]}$	0.23$_{[.19,.29]}$
gpt4-turbo$_{pt}$	0.85$_{[.82,.88]}$	0.85$_{[.83,.89]}$	0.85$_{[.83,.88]}$	0.86$_{[.82,.89]}$	**0.86**$_{[.82,.89]}$	0.86$_{[.83,.88]}$
o1-mini$_{pt}$	0.79$_{[.75,.82]}$	0.88$_{[.84,.92]}$	0.83$_{[.81,.86]}$	0.87$_{[.83,.91]}$	0.76$_{[.72,.80]}$	0.81$_{[.79,.85]}$
RoBERTa$_i$	0.85$_{[.82,.88]}$	0.93$_{[.89,.95]}$	**0.89**$_{[.87,.91]}$	**0.91**$_{[.89,.94]}$	0.84$_{[.80,.88]}$	**0.88**$_{[.85,.90]}$

Table 2: Effectiveness of incident detection, the scores with highest mean value are **bolded**. Evaluation on RoBERTa$_i$ on the *has-incident* class is computed by conflating hassle, uplift, and mix predictions in comparison with gold labels.

Model	Hassle			Uplift			Mix		
	Precision	*Recall*	F_1	*Precision*	*Recall*	F_1	*Precision*	*Recall*	F_1
					Step 2				
RoBERTa$_{ft}$	0.90$_{[.84,.95]}$	0.82$_{[.76,.87]}$	0.86$_{[.82,.90]}$	0.87$_{[.80,.94]}$	0.69$_{[.62,.76]}$	0.77$_{[.71,.82]}$	0.56$_{[.49,.65]}$	0.80$_{[.72,.88]}$	0.66$_{[.59,.74]}$
llama3$_{pt}$	0.91$_{[.85,.96]}$	0.74$_{[.67,.80]}$	0.82$_{[.77,.86]}$	0.74$_{[.65,.80]}$	0.92$_{[.87,.96]}$	0.82$_{[.74,.86]}$	0.60$_{[.48,.73]}$	0.57$_{[.43,.66]}$	0.58$_{[.49,.68]}$
llama3$_{ft}$	0.77$_{[.70,.84]}$	0.83$_{[.76,.89]}$	0.80$_{[.76,.84]}$	0.90$_{[.79,.98]}$	0.35$_{[.26,.43]}$	0.50$_{[.40,.59]}$	0.38$_{[.30,.45]}$	0.65$_{[.53,.75]}$	0.48$_{[.39,.55]}$
gemma2$_{pt}$	0.82$_{[.74,.90]}$	0.72$_{[.63,.79]}$	0.76$_{[.70,.81]}$	0.71$_{[.61,.78]}$	**0.94**$_{[.89,.97]}$	0.80$_{[.73,.85]}$	0.48$_{[.36,.62]}$	0.36$_{[.26,.47]}$	0.41$_{[.30,.52]}$
gpt4-turbo$_{pt}$	0.88$_{[.81,.93]}$	**0.91**$_{[.86,.95]}$	**0.89**$_{[.85,.93]}$	**0.89**$_{[.83,.95]}$	0.79$_{[.73,.85]}$	**0.84**$_{[.79,.88]}$	**0.70**$_{[.61,.77]}$	0.77$_{[.70,.85]}$	**0.73**$_{[.67,.80]}$
o1-mini$_{pt}$	**0.94**$_{[.90,.98]}$	0.79$_{[.72,.85]}$	0.86$_{[.81,.90]}$	**0.89**$_{[.82,.94]}$	0.74$_{[.66,.80]}$	0.81$_{[.74,.85]}$	0.59$_{[.50,.67]}$	**0.86**$_{[.79,.93]}$	0.70$_{[.62,.77]}$
					end-to-end				
RoBERTa$_i$	0.66$_{[.58,.73]}$	0.77$_{[.70,.83]}$	0.71$_{[.66,.77]}$	0.65$_{[.56,.73]}$	0.67$_{[.57,.76]}$	0.66$_{[.58,.73]}$	0.43$_{[.34,.55]}$	0.45$_{[.34,.57]}$	0.44$_{[.35,.53]}$
RoBERTa$_e$	**0.85**$_{[.79,.91]}$	**0.83**$_{[.77,.89]}$	**0.84**$_{[.79,.88]}$	**0.86**$_{[.79,.92]}$	**0.87**$_{[.81,.93]}$	**0.86**$_{[.81,.91]}$	**0.86**$_{[.79,.93]}$	**0.96**$_{[.93,1.0]}$	**0.91**$_{[.87,.95]}$

Table 3: Effectiveness of incident-triggered sentiment classification, the scores with the highest mean value are **bolded**. RoBERTa$_{ft}$ and other LLMs are evaluated on instances with hassle, uplift, or mix labels, excluding non-incident cases, i.e., Step 2 only. RoBERTa$_i$ and RoBERTa$_e$ are evaluated on all instances as an end-to-end HUD system.

by each model and ground-truth label is shown in Table 9, Appendix E. We found that RoBERTa made fewer errors (24%) than any of the four LLMs on these two kinds of instances.

To gain insights into the nature of these prediction errors, we performed a qualitative manual investigation. Starting with instances having ground-truth "*non-incident*" label, we find that the language expressed in such instances often involves rich narrative accounts of sudden mental breakdowns or retrospective evaluations of concrete life context. While the narratives of these posts may contain temporal or sentiment-bearing content, they still are not valid incidents. For example, post (a) in Table 4 describes a raised tiredness, with a time reference "*8am*" and spatial information "*The Bus stop*", but the post does not mention any incident that triggers the tired feeling. We observed that RoBERTa is more robust in distinguishing such challenging expressions while large LLMs, especially the two open-resource models, often incorrectly flag such expressions as incident. In contrast, the disclosure of a transient feeling without any elaboration of time reference, spatial information, or concrete life situation, (Example (b) in Table 4, fall in the category of "easy" case, where all or at least four models can correctly identify.

Turning to the ground-truth "*has-incident*" posts, we observe that expressions in challenging posts typically lack overt linguistic features that conventionally signal an incident affecting the post's author, namely: (1) a first-person pronoun (e.g., *I*), (2) a verb denoting an action, and (3) a time reference. For example, the post (c) in Table 4 implies the incident that the individual has officially become a dedicated fan of THE BOYZ (a K-pop group). However, in the absence of specifying who became a fan and a time reference to this sudden event, all five models misclassified it as expressing a non-incident. In contrast, as humans, we can still infer the incident disclosed in this message. To justify our assumption that the incident detection has an over-reliance on these illustrated linguistic features, we rephrased 20 such instances in our dataset, such

	Example
(a)	*8am. The Bus Stop. Let's hope I'll get less tired throughout the day. #Tired*
(b)	*Why am I so pathetic? #Sad*
(c)	*Officially stanned THE BOYZ! #Sunny*
(d)	*I just took 2 edibles... I might die tonight lol #Gummy*
(e)	*I got the tip of a needle stuck in my finger. It no longer hurts #Renewed*

Table 4: Examples of successful and failed incident detection. (a) A general emotional reflection incorrectly identified as incident; (b) an expression of emotion without explicit temporal, spatial, or situational details correctly identified as non-incident; (c) an uplift lacking explicit subject, verb, or temporal reference misclassified as non-incident; (d) a hassle disclosed with sarcastic tone incorrectly identified as an uplift; (e) an uplift expressed through strongly negative or negated lexical cues incorrectly identified as a hassle.

as by changing into "*I officially stanned THE BOYZ today! #Sunny*", and re-applied incident detection. The result shows RoBERTa and the two proprietary LLMs can correct their predictions on all 20 instances. This observation suggests that, although language models, especially LLMs, have proven strong capability in language understanding, they may still fail to account for implicitly conveyed incidents, as often happens in social media context.

Classification of Incident-Triggered Sentiment: Applying the same difficulty categorization, we identified 218 easy, 82 moderate, and 23 highly challenging posts. Model performances on moderate and highly challenging cases are summarized in Table 10, Appendix E. Notably, the two proprietary models (`gpt4-turbo` and `o1-mini`) made more errors than $\texttt{RoBERTa}_{ft}$. Interestingly, the open-source LLMs made fewer errors than the other models in identifying uplifts.

Expressions in challenging *hassle* posts often involve sarcasm (e.g., (d) in Table 4), which mislead all models to assign mixed or positive labels. This is a known challenge for sentiment analyzers. In HUD, we found errors induced by sarcastic expressions are mostly prevalent in open-source LLMs, followed by the two proprietary models and then $\texttt{RoBERTa}_{ft}$. Additionally, when posts contain both positive and negative expressions, but the incident itself triggers only positive sentiment, models are likely to label mixed valence, such as the Example (d) in Table 1, indicating a failure to link the sentiment triggered by the incident. Overall, $\texttt{RoBERTa}_{ft}$ makes the fewest such errors.

Expressions in challenging *uplift* posts exhibit the inverse pattern. When users express uplifts using metaphor or informal phrasing (e.g., "*cute shit*", "*it was a big ass hit :')*"), models often flag them as mixed or even negative example. We hypothesize that there might be stereotypical word associations in language models—for instance, terms like "*hurt*" or "*stuck*" are frequently linked with negativity (Chen et al., 2025), regardless of post content or the self-selected tags that actually signal positive response. $\texttt{RoBERTa}_{ft}$ specifically shows difficulty in handling negation, misclassifying posts like example (e) in Table 4 as hassle, when it actually indicates a positive feeling (an uplift), signalled by the tag "*#Renewed*".

Expressions in challenging *mixed* posts also feature sarcasm, metaphor, and negation, and include cases where one sentiment is dominant, e.g., most of the post content expresses hassles, but it also contains a few short phrases indicating an uplift. The open-source LLMs tend to follow the dominant sentiment cue(s), leading to biased predictions to either hassle or uplift. In contrast, $\texttt{RoBERTa}_{ft}$ and the proprietary models are more likely to capture the minor contrasting tone in the expression.

In summary, our qualitative analysis highlights the key challenges in HUD, including distinguishing between posts mentioning incidents and posts mentioning other aspects of life, and disentangling incident-triggered sentiment from other aspect-based sentiment. At the same time, HUD also struggles to handle sarcasm, metaphor, and negation, which are known challenges in existing sentiment analysis studies.

6 Psycholinguistic Analysis of Writing Styles of Vent vs. Other Text Resources

Psycholinguistic analysis are effective at measuring the writing style of text from a lexical-psychological perspective (Pennebaker et al., 2015). Given that many studies involving sensitive personal data are restricted from public disclosure, we propose an alternative strategy for offering insights into the linguistic characteristics of the data while upholding privacy commitments.

Hassles and uplifts are often embedded in emotionally expressive content, typically marking the onset of cognitive processes. We characterize the writing styles of such cognitive process using LIWC (Linguistic Inquiry and Word Count) (Pen-

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

Table 5: Average LIWC score related to cognitive process among various data sources.

nebaker et al., 2015). LIWC uses predefined word lists to compute the proportion of words in each post that fall into specific categories associated with cognitive processes. We apply LIWC to a sample of $20,689$ randomly selected posts from Vent, corresponding to all the posts of the users selected in Step 1 of our data selection. We compare the resulting psycholinguistic profiles to other well-documented corpora analyzed using LIWC2015 (Pennebaker et al., 2015), including general tweets, personal blogs, novels, diaries used in conventional psychological study (Pennebaker and Chung, 2007), New York Times articles, and samples of natural speech. The LIWC scores for the compared datasets (e.g., blogs, novels, speeches) are drawn from aggregate statistics reported by the LIWC developers (Pennebaker et al., 2015), without access to instance-level data. To account for statistical significance, we follow their practices by reporting the total number of instances underlying each dataset (Table 5). The amount of instances supporting the statistical comparison are sufficiently large (N > 6000) for most of the datasets, suggesting that the relative differences in each category of LIWC score between two data sources exceeding 0.5 will indicate statistical significance.

We observe that Vent posts have a relatively high-ratio use of cognitive process words (11.70) (Table 5). This aligns with the nature of Vent, where people share their momentary life experience. In particular, Vent scores higher than Tweets (9.96), which suggests that Vent posts include more reasoning and reflection of personal context. Vent exhibits particularly high expressions of *discrepancy* (1.94) and *tentative* (2.98) words. This 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 Vent users prefer to make distinctions between concepts or emotions, which may show various coping strategies. Although this comparison may not capture all stylistic dimensions, it still reveals that the linguistic profile of Vent posts closely mirrors the diaries commonly used in psychological research. This resemblance underscores the suitability of using LIWC for dataset characterization.

To the best of our knowledge, no publicly available resources exhibit a similar writing style or spontaneous data collection procedure as our datasets or diaries privately held by psychologist (Pennebaker and Chung, 2007) due to ethical and privacy concerns. Our aim in reporting the psycholinguistic characteristics of our data source is to facilitate future comparison.

7 Related Work

Researchers have utilized NLP techniques to tackle various mental health-related tasks. Some 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) expressions, or detect depressive disorders (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 emotion it elicited.

While LLMs have proven effective across many NLP tasks, their reliability for HUD remained unexamined. Prior work shows that BERT-style small language models (SLMs) outperform generative LLMs on sentiment classification in low-resource

settings (Barbieri et al., 2020; Bucher and Martini, 2024), particularly with contrastive adaptation (Tunstall et al., 2022). However, it is unclear whether such effectiveness applies to HUD.

8 Conclusion

We introduced a novel hassles and uplift detection (HUD) task that focuses on extracting specific concepts (hassles and uplifts, or mix, i.e., negative or positive responses to incidents) grounded in psychological theory. Through a series of analyses of existing language models and task adaptation strategies, we found that identifying hassles, uplifts and mixes of both still presents challenges to the current NLP approaches, and results are not yet reliable. In contrast, human annotators can effectively distinguish the sentiments triggered by incidents from general sentiment expressions. This clearly demonstrates the need for further work on the HUD task. Furthermore, we proposed an approach to overcome the common challenge of releasing sensitive mental health data alongside experimental results by reporting the psycholinguistic profile of text resources.

In the future, we will seek to improve the effectiveness of our HUD system. We will extend the evaluation to other publicly available datasets containing emotional expressive texts. We will extrinsically evaluate HUD on downstream tasks, such as on detecting "*moments of change*" in individuals' moods over time (Tsakalidis et al., 2022). Importantly, we will build on HUD to support research in emotion regulation and resilience.

Limitations

In this paper, we consider 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 about hassles but rather a post of pure emotional awareness.

We 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 social media platform. The data is also private due to its sensitivity (mental health) and potential risk of having identifiable information.

Ethical Concerns

We have obtained permission from the owners of the Vent platform at the time the data was shared with us, and ethics approval from our institution, CSIRO, Australia (approval: 217/23), to use and annotate 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 securely within the organization. We have used examples of the posts that are not identifiable. The annotators had no access to the post creators identities. The proprietary LLMs (`gpt4-turbo` and `o1-mini`) are hosted within our organization server, ensuring the privacy of processing user-sensitive data.

Potential Risks

The tested language models carry the risk of producing biased and potentially harmful predictions. 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.

Licenses of Artifacts

All scientific artifacts cited or utilized in this paper were employed in accordance with their respective license of use.

References

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

Jun Araki and Teruko Mitamura. 2018. Open-domain event detection using distant supervision. In *Proceedings of the 27th international conference on computational linguistics*, pages 878–891.

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.

Niall Bolger, Angelina Davis, and Eshkol Rafaeli. 2003. Diary methods: Capturing life as it is lived. *Annual Review of Psychology*, 54(1):579–616.

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, Sarvnaz Karimi, Diego Mollá, and Cécile Paris. 2025. To labor is not to suffer: Exploration of polarity association bias in llms for sentiment analysis. In *Proceedings of the fourteenth International Joint Conference on Natural Language Processing & fourth Asia-Pacific Chapter of the Association for Computational Linguistics*.

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.

Monique F Crane, Ben J Searle, Maria Kangas, and Y Nwiran. 2019. How resilience is strengthened by exposure to stressors: The systematic self-reflection model of resilience strengthening. *Anxiety, Stress, & Coping*, 32(1):1–17.

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.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392.

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*.

Bradley Efron. 1992. Bootstrap methods: another look at the jackknife. In *Breakthroughs in Statistics: Methodology and distribution*, pages 569–593. Springer.

Samantha L Falon, Maria Kangas, and Monique F Crane. 2021. The coping insights involved in strengthening resilience: The self-reflection and coping insight framework. *Anxiety, Stress, & Coping*, 34(6):734–750.

James J Gross. 1998. The emerging field of emotion regulation: An integrative review. *Review of general psychology*, 2(3):271–299.

James J Gross. 2015. Emotion regulation: Current status and future prospects. *Psychological inquiry*, 26(1):1–26.

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*.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177.

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.

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, Andreas Duenser, Maria Kangas, Diego Mollá-Aliod, and Cecile Paris. 2023. Message similarity as a proxy to repetitive thinking: Associations with non-suicidal self-injury and suicidal ideation on social media. *Computers in Human Behavior Reports*, 11:100320.

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.

Inez Myin-Germeys, Margreet Oorschot, Dina Collip, Johan Lataster, Philippe Delespaul, and Jim Van Os. 2009. Experience sampling research in psychopathology: opening the black box of daily life. *Psychological medicine*, 39(9):1533–1547.

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.

Ambreen Nazir, Yuan Rao, Lianwei Wu, and Ling Sun. 2020. Issues and challenges of aspect-based sentiment analysis: A comprehensive survey. *IEEE Transactions on Affective Computing*, 13(2):845–863.

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*.

Arthur A Stone, Stefan Schneider, and Joshua M Smyth. 2023. Evaluation of pressing issues in ecological momentary assessment. *Annual Review of Clinical Psychology*, 19(1):107–131.

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.

Elsbeth Turcan, Smaranda Muresan, and Kathleen McKeown. 2021. Emotion-infused models for explainable psychological stress detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2895–2909.

Michael Ungar. 2013. Resilience, trauma, context, and culture. *Trauma, violence, & abuse*, 14(3):255–266.

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/.

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.

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.

A Catalog of Daily Minor Incidents

The catalog shown in Table 6 is derived from (Kanner et al., 1981) and further curated by a clinical psychologist (co-author of this paper).

B Prompt Template for LLMs

The exact prompts used in our experiments are shown below. Each prompt was constructed using the chat template specific to its corresponding LLM. To assess the impact of prompt variation, we initially tested multiple prompt designs on a subset of 100 instances. We observed no substantial differences in performance across these variants, likely due to the use of in-context learning with the full set of training instances from each cross-validation fold. Consequently, we adopted the following two fixed prompts for all subsequent experiments. In contrast, we report the non-bootstrapping measure (Section 3.2) to present the 95% confidence interval.

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

- Prompt Template for **Incident-triggered Sentiment Classification**:
  `You are a psychologist and a text classifier. Does the incident described in the following text elicit of positive (1), negative (-1), or the mixture of both (0) sentiment? Requirement: Only answer 1 as positive, -1 as negative, 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 7.

D Statistics of Cross Validation Dataset

The statistics of five-fold cross validation set is shown in Table 8.

E The Error Rate on Challenging Posts

The error rate on challenging posts for incident detection is shown in Table 9 and for triggered sentiment classification is shown in Table 10.

F HUD Annotation Guideline

This guideline is shared with the annotators for their reference for the annotation of our dataset.

F.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.

F.2 Task description

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

F.3 Annotator Demographics

The four annotators of the HUD dataset are highly educated professionals based in English-speaking

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 6: Catalog of daily minor incidents with itemized sub-incidents.

params	value
GPU	NVIDIA RTX 3500 Ada
context size	512
temperature	0.01
quantization	4-bits
LoRA rank	32
LoRA alpha	32
target modules	lm_head
learning rate	$1e-5$
epoch	1

Table 7: The configurations for prompt-tuning or QLoRA fine-tuning of `Gemma2` and `Llama3`.

Fold	Incident Detection		Subj Feeling Detection		
	Incident	Non-incident	hassle	mix	uplift
$cv1_{eval}$	63	66	20	25	18
$cv2_{eval}$	73	57	36	23	14
$cv3_{eval}$	65	65	31	17	17
$cv4_{eval}$	60	71	21	20	19
$cv5_{eval}$	62	68	20	27	15
total	323	327	128	112	83

Table 8: The total count of validation instances per cross-validation fold.

Model	has-incident	non-incident	%Errors
$\texttt{RoBERTa}_{ft}$	19	46	24%
$\texttt{llama3}_{pt}$	23	211	84%
$\texttt{gemma2}_{pt}$	13	228	88%
$\texttt{gpt4-turbo}_{pt}$	34	47	30%
$\texttt{o1-mini}_{pt}$	27	75	37%

Table 9: The count of errors made on moderate and highly challenging posts by ground-truth label. The %Errors equals the proportion of errors made on the total amount of moderate and challenging posts.

Model	Hassle	Uplift	Mix	%Errors
$\texttt{RoBERTa}_{ft}$	16	22	14	30%
$\texttt{llama3}_{pt}$	31	9	33	70%
$\texttt{gemma2}_{pt}$	34	7	38	75%
$\texttt{gpt4-turbo}_{pt}$	21	16	13	48%
$\texttt{o1-mini}_{pt}$	25	24	10	56%

Table 10: The count of errors made on moderate and highly challenging posts by ground-truth label. The %Errors equals the proportion of errors made on the total amount of moderate and challenging posts.

countries. To maintain gender balance, the group includes two males and two females. Their native languages are Persian, French, Spanish, and Chinese.

F.4 Instructions

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

2. Decide whether the post conveys daily minor incident(s) that trigger sentiment in either positive, negative, or a mixture of both.

3. Select both *hassles* and *uplifts* label if you think a post describes both positive and negative incident-based sentiment; Do *NOT* make decision by only relying on the sentiment tone of the language.

4. Leave the cell *empty* if the text message is solely venting emotions or is a reflection on

generally existing circumstance without the explicit indication of occurrence.

5. Select *unknown* if you cannot decide the elicited emotion of the incident.

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

F.5 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 is specific, temporally bounded construct. In contrast, non-incident instances are likely to be solely containing emotional awareness or a description of a generally existing circumstance without clear indication on its occurrence. We incorporated a conservative annotation inference policy, i.e., only label "has-incident" to posts that describe tangible and objectively identifiable incident(s).

Examples of two has-incident instances:

(a) My back is killing :(#Floppy
(b) Visited a friend of mine in hospital, he's okay now #Relieved

Examples of two non-incident instances:

(c) 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. im truly alone again. feels like 2017 all over again :') #Sad

(d) 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

Specifically, while Example (c) may imply that an incident occurred in 2017, it may also indicate the self-reflection of someone's chronic negative experience. This expression lacks the mention of a tangible and objectively identifiable incident as compared to Example (a) & (b). Thus, instances like (c) will be treated as non-incident. Following suggestions by two psychological experts, annotators should avoid making over-implication on post content to minimize annotator bias and preserve annotation consistency.

F.6 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 or trauma (Ungar, 2013) are less frequent and can 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.

F.7 Definition of Incident-triggered Sentiment

Uplifts conveys experiences and conditions of acute daily incidents that have been appraised as positive or favorable to the post creator's well-being. **Hassle** conveys experiences and conditions of daily incidents that have been appraised as negative and harmful or threatening to the post creator's well-being. For the **mixed** cases, only posts with clearly identifiable co-occurring hassle and uplift components were annotated, while borderline cases, such as those with contrasting sentiment expression but without direct implication of incident trigger to both sides of the polarity will not be seen as a valid "mixed" case, to preserve label quality.

Below are examples of incident-triggered positive, negative, and several mix cases. Specially, the second uplift example shows a blend of positive and negative sentiment but only the positive is triggered by an objectively identifiable incident. Thus, it only expresses uplift instead of mix.

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

I'm feeling really down these days, but at least good to hear my mom is visiting me next month. #Recovering

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

Mix: *The exam yesterday was exhausting but I MADE THIS... I'VE DONE IT #Stormy*

Its going to be a nice day in the park for this family, leggo BUT shxx my stupid cousin is also coming. #Hyped

I am flying back to Oz tomorrow. I really enjoyed this vacation in Thailand. I am gonna so miss this trip. Oh no what to do #Bored

Had an awesome night out with friends! Finally got to unwind. But now I've got the world's worst hangover... #Regret

F.8 Handling Sarcasm, Irony, or Metaphor

Annotators must consider the user-selected emotional tag when inferring the sentiment triggered by the described incident. A post should be labeled as a hassle if it presents an incident with positive literal wording but is paired with a negative tag—for example, "*Wow, another wonderful day with double shifts! #Tired*". All annotations of incidents and their associated sentiments should be grounded in the literal content of the post. If an incident is conveyed metaphorically, it should still be annotated as an incident, as the veracity occurrence of the incident cannot be determined without access to the poster's background information.

G 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 11.

Model	Cost
`gpt4-turbo`	≈ 23.92
`o1-mini`	≈ 2.63

Table 11: 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.