Extraction of Texters' Explicit Emotion Expressions in Crisis Conversations

Greg Buda Crisis Text Line
Ignacio J. Tripodi Crisis Text Line
Kelly L. Zuromski Crisis Text Line
Margaret Meagher Crisis Text Line
Elizabeth A. Olson Crisis Text Line

Abstract

Understanding the emotions that individuals in crisis express is a clinically relevant goal. Here, we introduce an automated method for extracting present and past personal emotion expressions from text-based crisis conversations, enabling nuanced analyses of how these emotional profiles vary by age. We develop a three-tier emotion taxonomy and leverage both real conversation data and synthetic sentences to train a transformer-based model that captures contextual distinctions between true personal emotion expressions and other mentions. Our RoBERTa-based classifier outperforms both a regex baseline and a model trained only on real conversation data, achieving an F1 score of 0.856. Subsequent analysis of 338,924 crisis conversations shows that age is correlated with distinct patterns in emotional expressions. These findings underscore the clinical value of age-sensitive emotion analysis and constitute an initial step toward characterizing lexical variations across demographic groups.

1 Introduction

Negative emotions contribute to the development and maintenance of suicidal thoughts and behaviors (STBs), which is reflected in prominent theories of suicide. For instance, constructs indicating negative affective states, like psychological pain (Shneidman, 1993), hopelessness (Beck et al., 1975), and perceived burdensomeness (Joiner, 2005) are proposed to confer risk for STBs over time. Further, in the short-term, emotional disturbances are often present in the days and hours leading up to a suicide attempt. These crisis periods are characterized by intense, rapidly escalating negative affective states (Hendin et al., 2007; Rogers et al., 2017), which may include feelings of hopelessness, anger, anxiety, and agitation (Rudd et al., 2006). Because these 'warning signs' often precede suicidal behavior, understanding the emotions someone is experiencing can shed light on their imminent risk of suicide.

Effective emotion regulation (ER) strategies may differ across negative emotion categories; for example, sadness may be more easy to intentionally regulate than other emotions such as anger (Webb et al., 2012), and different ER strategies are maximally effective for different negative emotions (Quiñones-Camacho and Davis, 2020). Intense emotions resolve over different time courses across different emotion categories; fear naturally resolves more quickly than anger (Verduyn et al., 2009). Finally, reduced emotional vocabulary ('alexithymia') is strongly associated with suicidal ideation and attempts (Hemming et al., 2019).

Characterizing direct texter linguistic expressions of emotion in crisis conversations is therefore a clinically relevant goal. However, simple extraction of expressed emotion using a lexicon is inadequate, for multiple reasons. First, texters commonly discuss others' emotions during conversations ("I'm really stressed, my mom is so *mad* at me"). Second, texters frequently mention future or hypothetical emotions ("If my mom got mad at me, I would be *scared*").

Our objective is to automatically identify and extract present and past explicit personal emotion expressions in crisis conversations in order to examine how these emotion expression profiles vary by demographic characteristics, including age.

2 Literature Review

2.1 Related Work

A substantial body of work in emotion analysis has focused on *emotion detection and classification*, where the goal is to assign a label from a set of broad, discrete emotional categories (e.g., *anger, joy*) - often inferred from implicit expressions or textual cues. Most studies rely on taxonomies such as Ekman's (1992) six basic emo-

Findings of the Association for Computational Linguistics: ACL 2026, pages 27–44

tions or Plutchik's (1982) eight primary emotions as discrete target labels. Methods range from traditional CNN and recurrent neural networks (Sharker et al., 2022; Abdul-Mageed and Ungar, 2017), to prompt-tuning of large language models (Gao et al., 2024) and fine-tuning of BERT-based architectures (Rezapour, 2024). An alternative to the approach with discrete labels involves predicting real-valued intensity scores for individual emotion categories (Mohammad and Bravo-Marquez, 2017), or situating emotions within a multidimensional space such as Russell's (1980) arousal–valence model.

In contrast to the predictive nature of these studies, *emotion (span) extraction* typically appears as a subtask within multi-task frameworks. For example, the Emotion-Cause Pair Extraction (ECPE) task (Xia and Ding, 2019) identifies pairs of explicitly expressed emotions (e.g., "happy") and their corresponding causes (e.g., "received a flower"). Extensions to this task include extracting the pair of emotion and the resulting behavior or action (Sun et al., 2023), and the extraction of emotion-cause pairs as spans, followed by the classification of the emotion (Bi and Liu, 2020).

2.2 Our Contribution

Although emotion analysis has advanced rapidly in recent years, no current model or benchmark task is designed to reliably extract explicit emotion expressions in real-world mental health-related conversations. Our work addresses this gap through the following contributions:

1. *A clinically-vetted taxonomy of explicit emotion expressions.* Existing emotion lexicons often combine true emotional descriptors (e.g., happy, overwhelmed) with affective vocabulary (e.g. ouch!, smile), categorizing them into a fixed set of emotion labels. However, many of these associations are contextually ambiguous or nonsensical (Zad et al., 2021). Our taxonomy differs from the NRC Emotion Lexicon (Mohammad and Turney, 2013) in that it focuses on words directly denoting emotional states (e.g. irritated), while the NRC covers a broader range of associations. For example, the NRC assigns emotion labels to generic nouns (tree → anger), colors (white → anticipation) and places (mosque → anger). Unlike the NRC's flat mapping of words to basic emotions, our hierarchical structure groups lexemes into emotion families, capturing nuance and relationships between related emotions. Other existing formal hierarchies such as WordNet-Affect (Strapparava and Valitutti, 2004) focus on formal language and are not frequently updated, raising questions about their applicability for emotions related to current slang expressions of distress (Dhuliawala et al., 2016), particularly for younger texters.

2. *Extracting emotion expressions tied to the texter.* Knowing which emotion expression appears is only half the challenge - the other half is figuring out who is actually feeling that emotion. Most current methods ignore this distinction, treating all emotion mentions as equal regardless of context. A few studies have attempted to distinguish the experiencer, but these are limited to *classification* tasks rather than *span extraction*, and results show moderate performance (Wegge and Klinger, 2023; Kim and Vossen, 2021; Troiano et al., 2022). Lexicon-based methods augmented with syntactic heuristics (e.g. dependency parses) would fail to differentiate subject-experiencer psych words (e.g., "I love you") from object-experiencer psych words (e.g., "You scare me"), and they cannot identify the experiencer when it's implicit (e.g., "Feeling sad today").

3. *Filtering for real emotions only.* Simple lexicon-based approaches can misfire with negations (e.g., "not sad") and polysemy (e.g., "blue" as a color vs. sadness). To our knowledge, no existing annotation framework or context-sensitive extractor filters out non-actual states - hypothetical ("I would be happy if that happened"), future ("Maybe I'll be happy"), or desired ("I want to be happy") - ensuring only genuine emotional expressions are captured.

4. *Demographic-informed training.* The literature review by Plaza-del Arco et al. (2024) points out that most emotion-analysis research overlooks the demographic composition of its input data. We will specifically focus on age. Although comprehension of complex emotion vocabulary develops quickly during childhood and levels off after the age of 11 (Baron-Cohen et al., 2010), several aspects of emotional development continue to mature throughout adolescence. For instance, adolescents have a significantly larger emotion-specific vocabulary than younger children (Grosse and Streubel, 2024), but they often struggle more than both children and adults to differentiate between their emotional states (Nook et al., 2018). Age-related differences in emotional expression persist across adulthood: compared to younger adults, older adults use positive emotion expressions more frequently and negative emotion expressions less frequently (Löck-

enhoff et al., 2008). By incorporating age into a stratified sampling design and examining outcomes across different age groups, we aim to capture these developmental shifts in emotional language, with particular focus on children, adolescents and young adults, given the emergence of most mental health conditions during these developmental windows (Paus et al., 2008).

5. *A proof-of-concept application in mental health.* While NLP-driven emotion analysis holds promise for mental health care, its real-world deployment remains scarce. Prior work has demonstrated automated emotion *detection* (inference) of small sets of broad emotional categories in therapy sessions (Rasool et al., 2025; Tanana et al., 2021), and linked this to measures of symptom severity (Burkhardt et al., 2022). On the *extraction* front, Singh et al. (2023) applied emotion-cause extraction (ECE) to clinical conversations; however, in their approach, emotions were manually annotated and used as input to identify causes. To our knowledge, no study in the mental health domain has combined fully automated emotion-expression extraction with considerations for experiencer attribution, authenticity filtering, and demographic representativeness.

In sum, while our work draws on established NLP techniques, the primary methodological contribution of this work lies in the operationalization of a novel task: the simultaneous filtering of negations, hypotheticals, others' emotions, social pleasantries, meta-emotions, and diagnostic labels to isolate only genuine personal emotion expressions. This annotation framework, grounded in clinical expertise, fills a concrete gap left by existing emotion lexicons and extraction methods, which were not designed for the informal, demographically diverse language of crisis conversations.

3 Methods

3.1 Emotion Taxonomy

Our initial goal was to craft a taxonomy of words that explicitly describe emotional states, excluding implicit affective vocabulary contained in lexicons like the NRC emotion lexicon (Mohammad and Turney, 2013) or broader psycholinguistic tools such as LIWC (Pennebaker et al., 2001).

We began by collecting 1,207 candidate emotion expressions from diverse online sources (see Supplementary Table T1). After removing duplicates and normalizing to noun forms where possible, two clinical psychologist authors (EOA & KLZ) independently hand-reviewed each item to flag true emotion expressions. Prior to reconciliation through discussion, the annotators agreed in 83.6% of the cases, with a Cohen's kappa of 0.626. During the annotation, we identified certain groups of expressions as not actually describing emotions. These are listed in Table 1. Those that express emotions in at least one context were retained in the taxonomy. Ultimately, 442 words remained.

Shaver et al. (1987) argued that individuals naturally organize emotion concepts into a hierarchical structure, so hierarchical emotion taxonomies are intuitive and reasonable. Our emotion taxonomy consists of three levels. The middle tier - **emotion word families** - consists of these 442 manually collected and filtered entries. Above them are six **basic emotion** categories (*anger, disgust, fear, sadness, surprise*, and *enjoyment* in place of *happiness*) based on Ekman's classic six (Ekman, 1992). The two clinical psychologist authors independently assigned one of these six basic emotions to each of the 442 emotion families, with an agreement in 83.4% of the cases (Cohen's $\kappa = 0.787$).

The bottom tier of the hierarchy expands each family into its full set of morphological derivatives (verbs, nouns, adjectives, and adverbs), generated by GPT-4 and GPT-4 mini and then manually cleaned up (e.g., 'sad', 'sadden', 'sadness', 'sadder', 'saddest'). We call them **emotion lexemes**. Importantly, the lexicon includes a few multi-word emotion lexemes - typically phrasal-verb constructions - such as 'freaked out', 'fired up', and 'let down'. On the other hand, we do not include phrasal verbs in which the emotional meaning is already fully carried by the head word itself - for instance, 'scare' is included, while 'scare off' is not, because the particle 'off' modifies directionality but does not contribute additional emotional content.

Figure 1 offers a slice of the hierarchy, showing how expressions flow from basic categories down to specific lexemes.

3.2 Data

As a first step, an author (EAO) developed the initial annotation guidelines. Tagging was restricted to the words included in the emotion taxonomy described above. For polysemous or context-sensitive expressions, additional instructions were created to clarify when such words should be tagged as emotions. Table 2 outlines the key inclusion and

Category	Examples
Cognitive states	puzzled, questioning, interested, bored, undecided, confusion, conflicted
Interpersonal attitudes and dynamics	possessive, stifled, bullied, vulnerable, dependent, oversensitive, loyal
Social behavior descriptors	tactless, rude, assertive
Personality traits	extroverted, optimistic
Physical / somatic states	tired, hungry, nauseated, exhausted, uncomfortable
Psychiatric diagnosis or symptoms	suicidal
Generic negative affect	bad, stressed, upset, burned out, horrible
Generic positive affect	good, better

Table 1: Examples of expression groups not tagged as emotions.

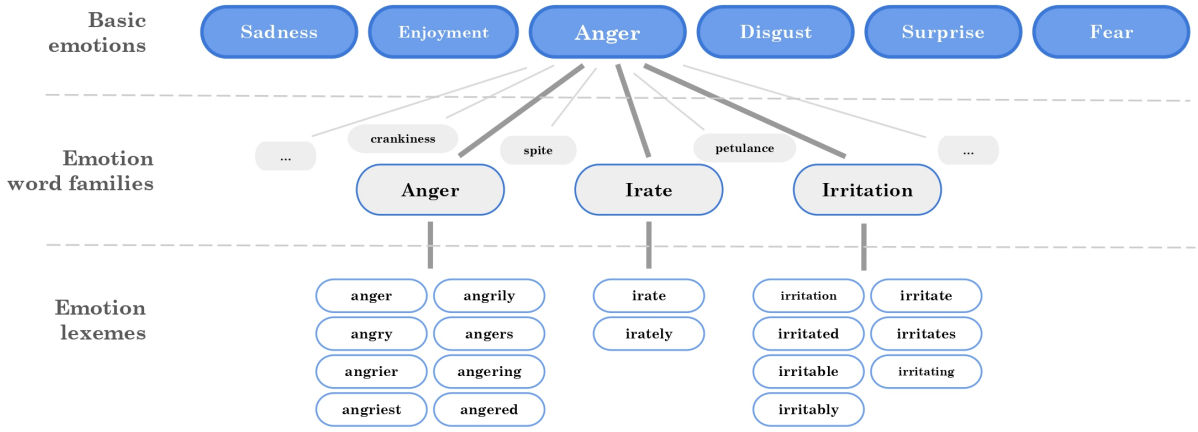


Figure 1: An illustrative subset of the emotion taxonomy.

exclusion criteria.

Supplementary Figure S1 illustrates the complete workflow for training, evaluation, and inference, as detailed in the subsections below. We trained our emotion-extraction model on a mix of real and synthetic data.

Our real corpus contains crisis conversation transcripts from Crisis Text Line, a non-profit organization that provides a fully text-based crisis counseling and intervention service. These conversations are made up of exchanges between texters in crisis and trained volunteer counselors. Personally identifiable information had been scrubbed from the conversations prior to data access. This research was evaluated by an independent IRB, which issued an exempt determination. Conversations for both the training and testing sets were sampled from 2018 to 2024, stratified by age, race, gender identity, and sexual orientation to ensure demographic coverage, with no individual texter appearing in more than one conversation across the entire dataset. We only used the texters' messages, and not the counselors'.

For training, we first segmented each turn into sentences using the *sat-12l-sm model* from the *wtpsplit* Python library (Minixhofer et al., 2023), which we chose for its punctuation-agnostic approach, crucial for the irregular punctuation seen in chat messages. The resulting training data set comprises 3066 sentences from 60 conversations. Then, two annotators - including EAO - marked emotion-lexeme spans (Cohen's κ = 0.629), resolving disagreements through discussion. Throughout this article, we use emotion *expressions* and emotion *spans* interchangeably to refer to these annotated units, which may consist of one or more words. In total, 296 spans were annotated in the training set.

In the test set, four annotators labeled emotion spans at the conversation level. Cohen's κ between annotator pairs ranged from 0.496 to 0.644 (see Supplementary Table T2 for details). Consensus was reached via discussion. A larger number of annotators was involved in the test set due to its higher complexity, as annotations were performed for the entire conversation transcript rather than for extracted messages. After reconciliation, we rolled mentions of expression variants (e.g. "angry", "anger", "angrily") up to their emotion family level (e.g. "anger"), resulting in 239 annotated emotion spans (218 emotion word families) in 60 test set conversations. The choice of training on lexemes at the sentence-level but evaluating on emotion families at the conversation level (1) keeps

Category	Excluded (not an emotion)	Included (emotion-tagged)
Negated emotions vs. real emotions	"I'd be lying if I said that I feel **satisfied**."	"I feel **satisfied** because I cleaned up my room."
Others' emotions vs. own emotions	"My classmate said he's feeling **down**."	"I am feeling **down** because of the bad weather."
Future vs. present or past emotions	"I'm going to be **pissed off** if I don't get a good grade."	"I am so **pissed off** right now."
Hypothetical vs. real emotions	"I would be **happy** if I had money."	"My plants make me really **happy**."
Social pleasantries vs. genuine emotions	"I **appreciate** your help today."	"I **appreciate** my mother so much; she's the best."
Physical vs. emotional states	"My knee **hurts** after the run."	"This divorce has **hurt** me."
Subject-experiencer psych verbs: others' vs. own emotions	"My teacher **hates** me."	"I **hate** my teacher."
Object-experiencer psych verbs: others' vs. own emotions	"I **scare** my father."	"My father **scares** me."
Diagnostic label vs. emotional experience	"My therapist diagnosed me with **anxiety** disorder."	"I'm feeling **anxious** about losing my keys."
Meta-emotions vs. direct emotions	"I **hate** feeling scared."	"I **hate** the whole system and everyone."

Table 2: Examples contrasting excluded vs. included emotion spans.
Examples are illustrative and are not drawn from actual conversations.

inputs within BERT variants' token limits, (2) prevents high-frequency lexemes from dominating in the evaluation, and (3) mitigates errors when spans or negations cross sentence boundaries.

We augmented our training set with 987 synthetic sentences (473 annotated spans) generated via OpenAI's GPT-4o, which we then manually proofread and annotated. We used two sets of prompts. Inspired by the attribute-driven prompting that Yu et al. (2023) proposed, we designed our first set of prompts as a static text scaffold containing placeholder slots, each of which was populated by random draws from a list of alternatives. The prompt template is illustrated in Supplementary Figure S2. We iterated over all 442 emotion word families in our taxonomy, producing one positive and one negative example each time. The prompts were designed to maximize content variety while ensuring that a range of writing styles - across different age groups and reading levels - appear in the resulting synthetic data. We embedded age and reading-level attributes so that our synthetic outputs mirror the Crisis Text Line texter population. The negative-example slot was filled with text options that each exemplify a different exclusion rubric from Table 2 (e.g., negated emotions, hypothetical states). The other set of prompts targeted further rubrics from Table 2, such as social pleasantries and physical vs. emotional states. These prompt sets appear in Supplementary Table S3. After discarding a few unrealistic examples that, in our judgment, reflected highly improbable sequences (e.g. "I wasn't *admired* by the scary movie, it was too boring."), we hand-annotated the remainder to confirm guideline adherence and to capture any additional emotion expressions. Supplementary Table T3 presents the distribution of annotated emotion expressions in the training set, including both annotated conversations and synthetic data.

3.3 Experimental setup

Our dataset - comprising both authentic crisis dialogues and synthetic sentences - was randomly divided into 75% for training and 25% for validation. We fine-tuned BERT-based models as token-level sequence labeling models for emotion span extraction using "Beginning, Inside, Outside" (BIO) tagging. This mirrors the annotators' task of identifying the exact textual span that expresses the texter's emotion. The predicted spans (lexemes) were then mapped to the corresponding emotion word families using our emotion taxonomy.

We conducted a full grid search over two architectures (roberta-base, bert-base-uncased), a fixed batch size of 128, dropout rates of [0.0, 0.1, 0.3], learning rates [1e-5, 5e-5], and epochs [5, 8].[1] After identifying the best hyperparameters by maximizing F1 in the validation set, we retrained on the entire training set and tested on our hold-out test conversations. Finally, we ran ablation experiments to compare performance against a simple regex baseline and a model trained only on partial data.

[1]All experiments ran on an Amazon EC2 g5.2xlarge instance (8 vCPUs, 32 GiB RAM, NVIDIA A10G GPU) using SpaCy configuration files. Key package versions: wtpsplit==2.1.5, spacy-transformers==1.3.8, spacy-annotator==2.1.4, spacy==3.7.2, transformers==4.47.1, torch==2.3.0

3.4 Inference

We applied our final model to all Crisis Text Line conversations from 2019–2024 in which the texter replied to the counselor's first message and reported their age in an optional post-conversation survey, and where the conversation was not flagged as a prank or test. This yielded 389,923 conversations for analysis.[2]

3.5 Emotion Profiles

For each conversation i, we counted the number of extracted spans for each of the six basic emotions - denoted by k - and converted these into relative frequencies summing to one:

$$\mathbf{p}_i = (p_{i,1}, \ldots, p_{i,6}), \quad \sum_{k=1}^{6} p_{i,k} = 1.$$

We then fit a Dirichlet regression via the `DirichletReg` R package (Maier, 2021), predicting $\mathbf{p}_i$ from age group (dummy-coded as 13 or younger, 14–17, 18–24, with 25+ as the reference) and conversation length (token count). Under the common parametrization, each concentration parameter $\alpha_{i,k}$ links to the covariate row $\mathbf{x}_i$ of design matrix $\mathbf{X}$ by

$$\log(\alpha_{i,k}) = \mathbf{x}_i^\top \boldsymbol{\beta}_k, \quad k = 1, \ldots, 6.$$

The expected proportion is

$$E[p_{i,k}] = \frac{\alpha_{i,k}}{\sum_{j=1}^{6} \alpha_{i,j}},$$

and the total concentration $\sum_{k=1}^{6} \alpha_{i,k}$ governs the precision (inverse variance) of the profile. Parameters $\boldsymbol{\beta}_k$ were estimated via maximum likelihood using a combined BFGS and Newton–Raphson routine.

4 Results

4.1 Model Performance and Ablation

The grid search on our validation split identified RoBERTa-base with a learning rate of 1e-5, 0.0 dropout, and 8 epochs as optimal, yielding an F1 of 0.860 on the validation set. For the conversation-only ablation, we ran an identical grid search restricted to the conversation data, which again selected the same hyperparameter configuration as the final model. For the synthetic-only ablation, we report two variants: (i) the final model's hyperparameters applied directly, and (ii) a grid search tuned specifically on synthetic data, which selected BERT-base-uncased with a learning rate of 5e-5, 0.0 dropout, and 8 epochs.

Table 3 reports a test-set F1 of 0.856 when evaluating at the conversation-level emotion-family granularity. The ablation results confirm that models trained on conversation data alone or synthetic data alone underperform the combined setting. Adding synthetic sentences to the crisis conversations drove recall upward by exposing the model to a wider variety of emotion expressions, while precision dipped only slightly. In contrast, a simple regular-expression baseline extracting all emotion lexemes from our taxonomy would suffer very low precision (and imperfect recall from uncorrected typos), since it cannot account for context.

Table 4 presents the model's performance across the six basic emotions. F1 is above average for Anger and Sadness but below average for Enjoyment. Due to the small sample sizes for Disgust and Surprise, performance on these categories should be interpreted with caution.

For additional examples of span prediction, Supplementary Figure S4 shows span predictions on real messages from Reddit[3]. These messages are accessible via the ConvoKit package (Chang et al., 2020).

4.2 Error analysis

In our qualitative review of misclassified cases, we identified a few recurring patterns. First, the model can conflate genuine emotions with routine social pleasantries, for instance interpreting "I'm happy you're okay" as authentic happiness rather than polite reassurance. Second, some errors arise from distinguishing clinical diagnosis from emotions, particularly in borderline contexts where even human raters may diverge (e.g., "I've been falling back into a depressive episode" (non-emotion: diagnostic label), "I'm pretty sure I've been depressed since childhood" (emotion)). Third, the model can miss words with typos, such as "thank full", or mishighlight "loved" when the texter's intended word is "moved". While it correctly handles some misspellings (e.g., "I feel anrgy about it"), additional normalization is needed to reliably map such variants to the correct emotion labels in our taxonomy.

[2] Inference was performed in parallel using PySpark on an Amazon EC2 i3.4xlarge instance (16 vCPUs, 122 GiB RAM).

[3] https://www.reddit.com

Model	F1	Precision	Recall
Regular expressions	0.470	0.310	0.975
Conversation-only (re-tuned; params match final	0.824	0.880	0.775
Synthetic-only (final model params)	0.609	0.858	0.472
Synthetic-only (re-tuned for synthetic)	0.659	0.854	0.537
Conversation + Synthetic (final model)	**0.856**	0.868	0.844

Table 3: Model performance metrics in the test set.

Basic emotion	F1	Precision	Recall	Support (TP + FN)
Sadness	0.881	0.881	0.881	84
Fear	0.849	0.894	0.808	73
Enjoyment	0.767	0.757	0.778	36
Anger	0.900	0.900	0.900	20
Disgust	1.000	1.000	1.000	2
Surprise	1.000	1.000	1.000	3

Table 4: Performance metrics in the test set by basic emotions.

4.3 Descriptive statistics: emotions in crisis conversations

We filtered our 389,923 inferred conversations down to 338,924 by removing multiple conversations from the same texter and keeping only the first conversation from each - this avoids dependencies that would require nested models. Of these, 298,673 conversations (88.1%) contained at least one extracted emotion span, and 58.7% contained three or more spans. The full span-count distribution is shown in Supplementary Figure S5.

Texters tended to be young, with 42.3% under 18. The most frequent basic emotions (tier 1) were Sadness (62%) and Fear (57%). In 21.8% of all conversations, counselors flagged suicide discussion in the post-conversation survey; this rate is higher amoung younger texters (30.3% for age 13 or younger; 24% for ages 14-17). Full distributions of basic emotions and age groups in the inference set are provided in Supplementary Tables T4 and T5.

4.4 Usage of emotion word families across age groups

We counted the number of spans corresponding to each emotion word family - the middle layer of our taxonomy - and plotted the relative frequencies of the top five emotion word families for each basic emotion separately (Figure 2; Supplementary Figure S6). Age-group comparisons show that "scared/fear" dominates Fear expressions among younger texters, while older texters additionally employ "anxious/anxiousness" and "afraid". Usage of the "hate" family declines with age, and older texters exhibit a broader Anger vocabulary, including "frustration/frustrated" and "anger/angry". Readers interested in the family-level distributions of the other basic emotions can refer to Supplementary Figure S6.

4.5 Emotion profiles: age differences

We fitted a Dirichlet regression to predict the composition (relative frequencies) of the six basic emotions as a function of age and conversation length. Table 5 presents the estimated beta coefficients. Positive coefficients ($\beta_k > 0$) correspond to a higher concentration parameter for that emotion - and thus to higher expected proportion - while negative coefficients ($\beta_k < 0$) correspond to lower expected proportions. All coefficients for *word count* are positive, so longer texts raise every α_k, yielding a sharper (more certain) emotion profile (see equations in Methods). However, because Disgust and Surprise have the smallest coefficients, their relative proportions shrink in longer messages. As for age covariates (baseline: Age 25+), younger groups express more Enjoyment and Anger, whereas older groups express more Fear, Sadness and Surprise. Figure 3 displays the age-specific emotion profiles implied by the estimated Dirichlet regression coefficients.

To test whether age-related differences were confounded by the higher prevalence of suicidality among younger texters in the sample, we refitted the model with an additional binary covariate indicating whether the conversation was flagged by counselors as involving suicide. As shown in Supplementary Table T6, the results remained stable, suggesting no confounding effect.

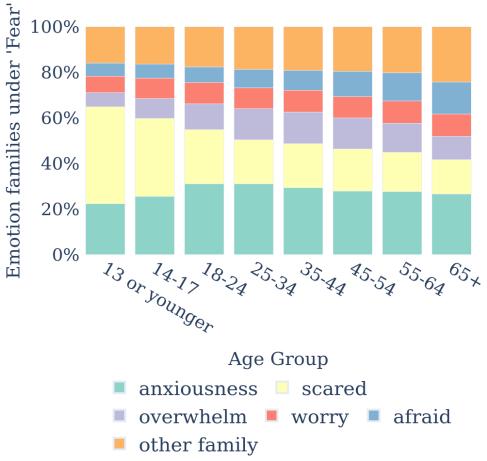
5 Discussion

We fine-tuned a RoBERTa-base model on both real crisis conversation data and purpose-built synthetic examples to identify only explicitly stated, genuine emotion expressions in texter messages. By leveraging context-sensitive representations rather than a simple lexicon lookup, our approach achieves a significant performance gain (F1 = 0.856 vs. 0.470). Furthermore, because manually annotating sparse emotion spans in real conversations is laborious - and most messages contain no emotion expressions

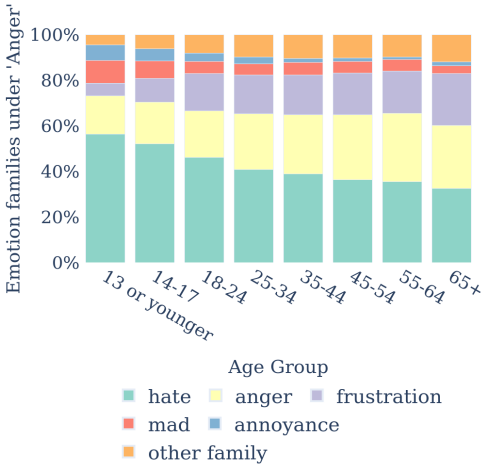
Predictor	Anger	Disgust	Enjoyment	Fear	Sadness	Surprise
Intercept	-2.427^{***}	-2.710^{***}	-2.291^{***}	-1.808^{***}	-1.663^{***}	-2.706^{***}
Age 13 or younger	0.039^{***}	-0.009	0.126^{***}	-0.172^{***}	-0.178^{***}	-0.018^{**}
Age 14–17	0.026^{***}	-0.004	0.107^{***}	-0.108^{***}	-0.127^{***}	-0.015^{**}
Age 18–24	0.003	0.003	0.060^{***}	0.020^{***}	-0.031^{***}	-0.006
Word Count	0.175^{***}	0.035^{***}	0.252^{***}	0.244^{***}	0.231^{***}	0.034^{***}

Significance codes: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Table 5: Dirichlet regression beta coefficients by basic emotion



(a) Fear



(b) Anger

Figure 2: Relative frequencies of top emotion word families across age groups under supercategories 'Fear' and 'Anger'.

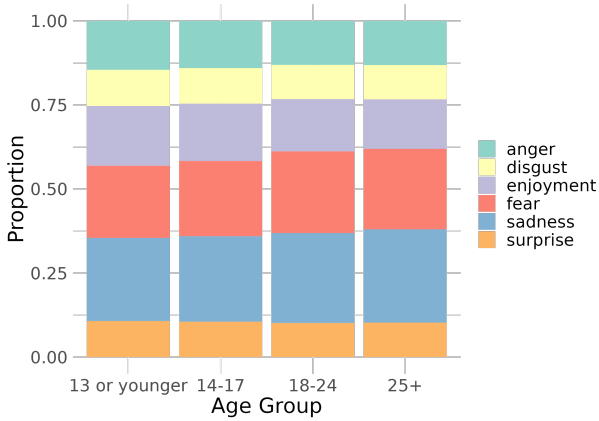


Figure 3: Predicted emotion profiles by age group, based on Dirichlet regression (holding word count constant at the sample median value).

- we supplemented our training data with synthetic sentences. This augmentation exposed the model to infrequent emotion expressions and diverse linguistic styles, substantially improving recall with only a modest impact on precision.

There are several possible interpretations of our findings that explicit emotion-expression profiles differ across age bands. First, it is possible that the situational factors that contribute to emotional crises differ across age groups, contributing to these different profiles. Second, developmental differences in emotion comprehension and abstraction may result in the emergence of different linguistic expressions of distress; prior work has shown that the level of abstraction in emotional expression continues to mature throughout adolescence (Nook et al., 2020). Additionally, it is possible that different emotional experiences may predominate in response to even comparable stressful events across age bands. Further work would be needed to differentiate these possibilities.

Our results carry some preliminary clinical implications. Younger texters showed higher usage of anger-related expressions and lower usage of sadness-related expressions than older texters. This may have important clinical implications for crisis intervention in this population, particularly since anger may be more difficult to intentionally regulate (Webb et al., 2012). Particular crisis counselor behaviors are associated with conversation outcomes (Buda et al., 2024); future work could focus on identifying conversational deescalation strategies that are particularly effective for texters

with particular emotion profiles. In addition, explicit emotion-expression profiles could serve as covariates or early indicators in suicide-risk detection models, contributing to more targeted and responsive crisis intervention and prevention systems.

As an additional finding, we observed that age groups often use different words to express the same or similar emotions (Figure 2; Supplementary Figures S6a–S6d). This constitutes an initial step toward characterizing lexical variation across demographic groups and suggests several avenues for future work. For example, researchers could examine differences in (i) intensity markers, (ii) action- versus state-oriented expressions (verbs vs. adjectives), and (iii) syntactic framing such as active vs. passive constructions (e.g., "Someone scares me" vs. "I scared someone"). Understanding these nuances would inform dataset construction and sampling strategies designed to reduce demographic-related bias in model training and evaluation.

6 Limitations

Some limitations of this work open avenues for refinement in future research.

Training and test set size: The number of annotated spans in the training and test sets is limited. Identifying 296 spans in the training set required manually reviewing 3,066 sentences, most of which contained no emotion-related expressions. Additional annotation in the test set - especially for less frequent basic emotions and emotion families - would enable more reliable conclusions at a granular level. Random sampling alone is unlikely to yield enough examples of rare expressions, so targeted strategies would be needed.

Typos and misspellings: Although misspellings and typographical errors were relatively rare in our training and test sets, we did not implement any specialized normalization technique; thus, these orthographic errors did not map to any emotion word family categories. Implementing dedicated spelling-correction algorithms could both improve extraction accuracy and reduce demographic skew.

Application in other domains: Supplementary Table S4 illustrates correct model predictions on Reddit messages, but a formal evaluation would be required to rigorously assess its generalizability to other conversational domains.

Further demographic dimensions: While our study focused exclusively on age as a demographic variable, other factors - such as gender or cultural background - could influence the usage of emotion expressions. Future synthetic-data generation and post-hoc analyses should integrate these dimensions to ensure broader representativeness.

Broad conversational context: Finally, our model is trained on individual texter sentences, even though we evaluate performance at the conversation level to capture broader context. Emotion recognition in conversation (ERC) frameworks demonstrate that leveraging context and inter-speaker dependency can enhance the inference of latent emotions (Fu et al., 2023). Although span-level extraction of explicit emotion expressions may be less reliant on conversational context than emotion inference tasks, scenarios such as a counselor's prompt ("How are you feeling?") followed by a simple reply ("Sad") illustrate the potential value of contextual cues.

7 Ethical Considerations

Conversation data are securely stored on a cloud platform. All personally identifiable information was removed from the conversations before data access. Demographic data are collected via an optional post-conversation survey. Although data are deidentified, they are still considered highly sensitive. For this reason, we cannot share the annotated dataset. To maximize reproducibility, we instead provide the emotion lexicon and include several out-of-sample examples and prompting strategies throughout the manuscript.

We present emotions extracted from texters of diverse ages, races, ethnicities, gender identities, and sexual orientations, which is a strength. However, a formal bias evaluation to assess for differential model performance across demographic groups has not yet been performed, because of the extensive hand-coding required for validation in this project. We acknowledge that the synthetic messages generated with LLMs could also introduce bias.

It is important to weigh the benefits versus the possible harms of conducting NLP research using sensitive datasets, and to take steps to minimize the risk of harm. We believe that this work opens new avenues to understand the role of emotion in mental health crises and may ultimately benefit texters. For example, this work provides a foundation for identifying effective deescalation strategies in people expressing emotions associated with increased suicide risk, like anger. These benefits are weighed

against potential risks, including risks related to bias and privacy risks. The data are handled by an in-house team trained in data protection and stored with strong safeguards. We believe that the risks of working with this dataset are balanced by the potential benefits of understanding the role of emotion in mental health crises.

Acknowledgments

This work was made possible by generous support to Crisis Text Line from the Jensen and Lori Huang Foundation.

References

Muhammad Abdul-Mageed and Lyle Ungar. 2017. EmoNet: Fine-Grained Emotion Detection with Gated Recurrent Neural Networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 718–728, Vancouver, Canada. Association for Computational Linguistics.

Simon Baron-Cohen, Ofer Golan, Sally Wheelwright, Yael Granader, and Jacqueline Hill. 2010. Emotion Word Comprehension from 4 to 16 Years Old: A Developmental Survey. *Frontiers in Evolutionary Neuroscience*, 2.

A. T. Beck, M. Kovacs, and A. Weissman. 1975. Hopelessness and suicidal behavior. An overview. *JAMA*, 234(11):1146–1149.

Hongliang Bi and Pengyuan Liu. 2020. ECSP: A New Task for Emotion-Cause Span-Pair Extraction and Classification. *arXiv preprint*. ArXiv:2003.03507.

Greg Buda, Ignacio J. Tripodi, Margaret Meagher, and Elizabeth A. Olson. 2024. Crisis counselor language and perceived genuine concern in crisis conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7149–7160, Miami, Florida, USA. Association for Computational Linguistics.

Hannah Burkhardt, Michael Pullmann, Thomas Hull, Patricia Areán, and Trevor Cohen. 2022. Comparing emotion feature extraction approaches for predicting depression and anxiety. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 105–115, Seattle, USA. Association for Computational Linguistics.

Jonathan P. Chang, Caleb Chiam, Liye Fu, Andrew Z. Wang, Justine Zhang, and Cristian Danescu-Niculescu-Mizil. 2020. ConvoKit: A Toolkit for the Analysis of Conversations. *arXiv preprint*. ArXiv:2005.04246 [cs].

Shehzaad Dhuliawala, Diptesh Kanojia, and Pushpak Bhattacharyya. 2016. SlangNet: A WordNet like resource for English Slang. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4329–4332, Portorož, Slovenia. European Language Resources Association (ELRA).

Paul Ekman. 1992. Facial Expressions of Emotion: New Findings, New Questions. *Psychological Science*, 3(1):34–38.

Yao Fu, Shaoyang Yuan, Chi Zhang, and Juan Cao. 2023. Emotion Recognition in Conversations: A Survey Focusing on Context, Speaker Dependencies, and Fusion Methods. *Electronics*, 12(22):4714.

Qingqing Gao, Jiuxin Cao, Biwei Cao, Xin Guan, and Bo Liu. 2024. CEPT: A Contrast-Enhanced Prompt-Tuning Framework for Emotion Recognition in Conversation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2947–2957, Torino, Italia. ELRA and ICCL.

Gerlind Grosse and Berit Streubel. 2024. Emotion-specific vocabulary and its relation to emotion understanding in children and adolescents. *Cognition & Emotion*, pages 1–10.

Laura Hemming, Peter Taylor, Gillian Haddock, Jennifer Shaw, and Daniel Pratt. 2019. A systematic review and meta-analysis of the association between alexithymia and suicide ideation and behaviour. *Journal of Affective Disorders*, 254:34–48.

Herbert Hendin, John T. Maltsberger, and Katalin Szanto. 2007. The role of intense affective states in signaling a suicide crisis. *The Journal of Nervous and Mental Disease*, 195(5):363–368.

Thomas Joiner. 2005. *Why people die by suicide*. Why people die by suicide. Harvard University Press, Cambridge, MA, US.

Taewoon Kim and Piek Vossen. 2021. EmoBERTa: Speaker-Aware Emotion Recognition in Conversation with RoBERTa. *arXiv preprint*. ArXiv:2108.12009.

Corinna E. Löckenhoff, Paul T. Costa, and Richard D. Lane. 2008. Age differences in descriptions of emotional experiences in oneself and others. *The Journals of Gerontology. Series B, Psychological Sciences and Social Sciences*, 63(2):P92–99.

Marco J. Maier. 2021. Dirichletreg: Dirichlet regression. R package version 0.7-1, available on GitHub.

Benjamin Minixhofer, Jonas Pfeiffer, and Ivan Vulić. 2023. Where's the Point? Self-Supervised Multilingual Punctuation-Agnostic Sentence Segmentation. *arXiv preprint*. ArXiv:2305.18893.

Saif Mohammad and Felipe Bravo-Marquez. 2017. Emotion Intensities in Tweets. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 65–77, Vancouver, Canada. Association for Computational Linguistics.

Saif M Mohammad and Peter D Turney. 2013. NRC Emotion Lexicon.

Erik C. Nook, Stephanie F. Sasse, Hilary K. Lambert, Katie A. McLaughlin, and Leah H. Somerville. 2018. The Nonlinear Development of Emotion Differentiation: Granular Emotional Experience Is Low in Adolescence. *Psychological Science*, 29(8):1346–1357.

Erik C. Nook, Caitlin M. Stavish, Stephanie F. Sasse, Hilary K. Lambert, Patrick Mair, Katie A. McLaughlin, and Leah H. Somerville. 2020. Charting the development of emotion comprehension and abstraction from childhood to adulthood using observer-rated and linguistic measures. *Emotion (Washington, D.C.)*, 20(5):773–792.

Tomáš Paus, Matcheri Keshavan, and Jay N. Giedd. 2008. Why do many psychiatric disorders emerge during adolescence? *Nature Reviews Neuroscience*, 9(12):947–957.

J.W. Pennebaker, M.E. Francis, and R.J. Booth. 2001. Linguistic inquiry and word count: LIWC. *Mahwah, NJ: Lawrence Erlbaum Associates*.

Flor Miriam Plaza-del Arco, Alba Curry, Amanda Cercas Curry, and Dirk Hovy. 2024. Emotion Analysis in NLP: Trends, Gaps and Roadmap for Future Directions. *arXiv preprint*. ArXiv:2403.01222.

Robert Plutchik. 1982. A psychoevolutionary theory of emotions. *Social Science Information*, 21(4-5):529–553.

Laura E. Quiñones-Camacho and Elizabeth L. Davis. 2020. Children's awareness of the context-appropriate nature of emotion regulation strategies across emotions. *Cognition & emotion*, 34(5):977–985.

Abdur Rasool, Saba Aslam, Naeem Hussain, Sharjeel Imtiaz, and Waqar Riaz. 2025. nBERT: Harnessing NLP for Emotion Recognition in Psychotherapy to Transform Mental Health Care. *Information*, 16(4):301.

Mahdi Rezapour. 2024. Emotion Detection with Transformers: A Comparative Study. *arXiv preprint*. ArXiv:2403.15454.

Megan L. Rogers, Igor Galynker, Zimri Yaseen, Kayla DeFazio, and Thomas E. Joiner. 2017. An overview and comparison of two proposed suicide-specific diagnoses: Acute suicidal affective disturbance and suicide crisis syndrome. *Psychiatric Annals*, 47(8):416–420.

M. David Rudd, Alan L. Berman, Thomas E. Joiner, Matthew K. Nock, Morton M. Silverman, Michael Mandrusiak, Kimberly Van Orden, and Tracy Witte. 2006. Warning signs for suicide: theory, research, and clinical applications. *Suicide & Life-Threatening Behavior*, 36(3):255–262.

James A. Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178.

Arman Sharker, Mohammad Abdur Rafi Farhab, Tahrima Akter Tamanna, Umma Rumman, Md. Tanvir Rouf Shawon, and Nibir Chandra Mandal. 2022. A Cross-Corpus Deep Learning Approach to Social Media Emotion Classification. In *2022 4th International Conference on Sustainable Technologies for Industry 4.0 (STI)*, pages 1–6.

Phillip Shaver, Judith Schwartz, Donald Kirson, and Cary O'Connor. 1987. Emotion knowledge: Further exploration of a prototype approach. *Journal of Personality and Social Psychology*, 52(6):1061–1086. Place: US Publisher: American Psychological Association.

Edwin S. Shneidman. 1993. *Suicide as psychache: A clinical approach to self-destructive behavior*. Suicide as psychache: A clinical approach to self-destructive behavior. Jason Aronson, Lanham, MD, US.

Gopendra Vikram Singh, Soumitra Ghosh, Asif Ekbal, and Pushpak Bhattacharyya. 2023. DeCoDE: Detection of Cognitive Distortion and Emotion Cause Extraction in Clinical Conversations. In *Advances in Information Retrieval*, pages 156–171, Cham. Springer Nature Switzerland.

Carlo Strapparava and Alessandro Valitutti. 2004. WordNet-Affect: an affective extension of WordNet. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Yawei Sun, Saike He, Xu Han, and Ruihua Zhang. 2023. A New Model for Emotion-Driven Behavior Extraction from Text. *Applied Sciences*, 13(15):8700.

Michael J. Tanana, Christina S. Soma, Patty B. Kuo, Nicolas M. Bertagnolli, Aaron Dembe, Brian T. Pace, Vivek Srikumar, David C. Atkins, and Zac E. Imel. 2021. How do you feel? Using Natural language processing to automatically rate emotion in psychotherapy. *Behavior research methods*, 53(5):2069–2082.

Enrica Troiano, Laura Ana Maria Oberlaender, Maximilian Wegge, and Roman Klinger. 2022. x-enVENT: A Corpus of Event Descriptions with Experiencer-specific Emotion and Appraisal Annotations. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1365–1375, Marseille, France. European Language Resources Association.

Philippe Verduyn, Ellen Delvaux, Hermina Van Coillie, Francis Tuerlinckx, and Iven Van Mechelen. 2009. Predicting the duration of emotional experience: Two experience sampling studies. *Emotion*, 9(1):83–91. Publisher: American Psychological Association.

Thomas L. Webb, Eleanor Miles, and Paschal Sheeran. 2012. Dealing with feeling: A meta-analysis of the

effectiveness of strategies derived from the process model of emotion regulation. *Psychological Bulletin*, 138(4):775–808. Publisher: American Psychological Association.

Maximilian Wegge and Roman Klinger. 2023. Automatic emotion experiencer recognition. In *Proceedings of the 3rd Workshop on Computational Linguistics for the Political and Social Sciences*, pages 1–7, Ingolstadt, Germany. Association for Computational Linguistics.

Rui Xia and Zixiang Ding. 2019. Emotion-Cause Pair Extraction: A New Task to Emotion Analysis in Texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1003–1012, Florence, Italy. Association for Computational Linguistics.

Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. 2023. Large Language Model as Attributed Training Data Generator: A Tale of Diversity and Bias. *arXiv preprint*. ArXiv:2306.15895.

Samira Zad, Joshuan Jimenez, and Mark Finlayson. 2021. Hell Hath No Fury? Correcting Bias in the NRC Emotion Lexicon. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 102–113, Online. Association for Computational Linguistics.

Supplementary Materials

The emotion taxonomy and annotated synthetic data are publicly available at: https://github.com/gbuda-ctl/Extraction-of-Texters-Explicit-Emotion-Expressions-in-Crisis-Conversations

Source Name / Organization	URL
SOCOM Feeling Words	https://www.socom.mil/POTFF/Pages/Feeling%20words.aspx
Notre Dame Academy Feelings Word List	https://www.ndapandas.org/wp-content/uploads/archive/Documents/News/FeelingsWordList.pdf
Kansas State University Counseling Services	https://www.k-state.edu/counseling/services/resources/self_help/feelingwordlist.html
APA Helping Skills Emotion Words Checklist	https://www.apa.org/pubs/books/supplemental/Helping-Skills-Fifth/EmotionWordsChecklist.pdf
Hoffman Institute Feelings and Sensations List	https://www.hoffmaninstitute.org/wp-content/uploads/Practices-FeelingsSensations.pdf
Berkeley Well-Being Institute Printable List of Emotions	https://www.berkeleywellbeing.com/uploads/1/9/4/8/19481349/printable-list-of-emotions.pdf
Karla McLaren Emotional Vocabulary List	https://karlamclaren.com/emotional-vocabulary-page/
The Work – Emotions List by Byron Katie	https://thework.com/wp-content/uploads/2019/02/Emotions_List_Ltr.pdf
Feelings PDF (SquareSpace site)	https://static1.squarespace.com/static/5ca4f4d1523958120344a27f/t/5cdc22080852297d846864f2/1557930505040/feelings.pdf
Tom Drummond Emotion Vocabulary	https://tomdrummond.com/leading-and-caring-for-children/emotion-vocabulary/

Table T1: Sources consulted for emotion vocabulary compilation (accessed in 2024)

	Annotator 2	Annotator 3	Annotator 4
Annotator 1	0.601	0.644	0.516
Annotator 2		0.630	0.496
Annotator 3			0.560

Table T2: Pairwise Cohen's κ between annotators on the test set. Each value was computed by considering all word families from our emotion lexicon that were present in a texters' messages, classifying each as emotion vs. non-emotion, and calculating Cohen's κ between these binary arrays.

Basic emotion	Number of annotated emotion spans in the training set
Sadness	227
Fear	156
Enjoyment	210
Anger	118
Disgust	21
Surprise	32

Table T3: Distribution of the annotated basic emotions in the training set (including conversations and synthetic data).

Basic emotion	% of Conversations in Inference Set
Sadness	62.3%
Fear	57.0%
Enjoyment	38.3%
Anger	26.7%
Disgust	1.4%
Surprise	1.2%

Table T4: Distribution of basic emotions across conversations in inference set.

Age group	% of Texters in Inference Set
13 or younger	13.3%
14–17	29.0%
18–24	25.1%
25+	32.6%

Table T5: Age distribution of texters in inference set.

Predictor	Anger	Disgust	Enjoyment	Fear	Sadness	Surprise
Intercept	-2.427^{***}	-2.710^{***}	-2.287^{***}	-1.795^{***}	-1.681^{***}	-2.705^{***}
Age 13 or younger	0.0389^{***}	-0.0088	0.1287^{***}	-0.1645^{***}	-0.1963^{***}	-0.0178^{**}
Age 14–17	0.0260^{***}	-0.0039	0.1082^{***}	-0.1039^{***}	-0.1368^{***}	-0.0145^{**}
Age 18–24	0.0026	0.0028	0.0605^{***}	0.0218^{***}	-0.0357^{***}	-0.0057
Word Count	0.1747^{***}	0.0352^{***}	0.2525^{***}	0.2447^{***}	0.2297^{***}	0.0338^{***}
Suicide Flag	0.0022	0.0002	-0.0245^{***}	-0.0700^{***}	0.1180^{***}	-0.0019

Significance codes: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Table T6: Dirichlet regression beta coefficients by basic emotion, with suicide flag as a covariate.

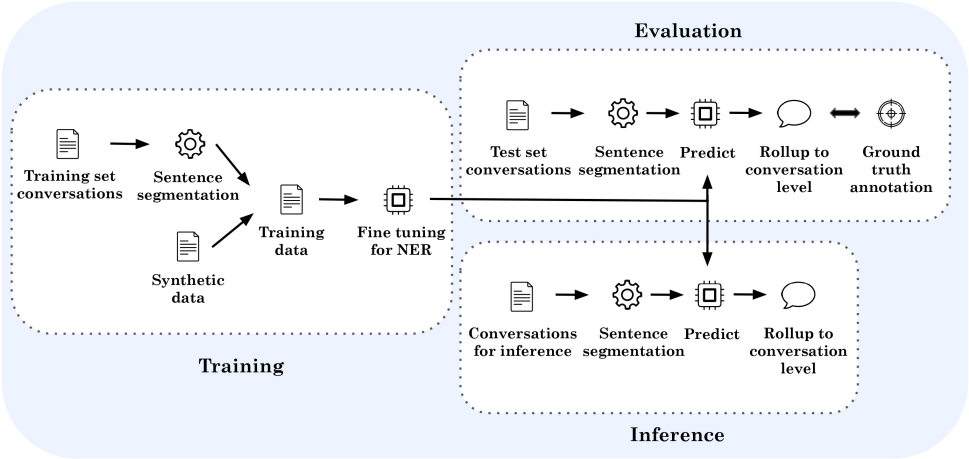
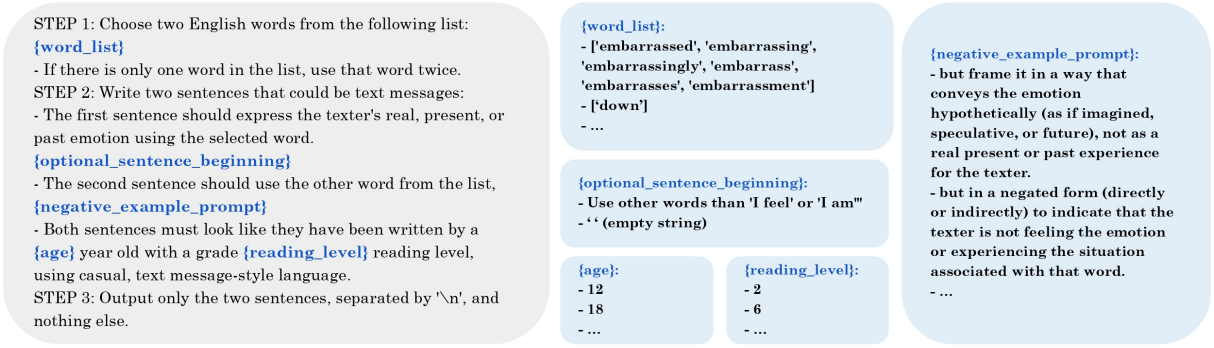


Figure S1: Overview of the full pipeline, including training, evaluation, and inference stages.



Slot	All options
{word_list}	All emotions families in our emotion taxonomy
{optional_sentence_beginning}	Use other words than 'I feel' or 'I am' *empty string*
{negative_example_prompt}	...but in a negated form (directly or indirectly) to indicate that the texter is not feeling the emotion or experiencing the situation associated with that word. ...but frame it in a way that conveys the emotion hypothetically (as if imagined, speculative, or future), not as a real present or past experience for the texter. ...but phrase it to convey the emotion as an obligation or expectation, rather than a real past/present experience. ...but frame it as part of a question or a question from someone else, without suggesting that this is the texter's actual personal experience. ...but it should reflect a third person's emotion, not the texter's.
{age}	8, 10, 12, 14, 16, 18, 20, 22, 24, 30, 40, 50, 60, 70, 80
{reading_level}	2, 4, 6, 8

Figure S2: Attributed prompt template for synthetic data generation.

Further prompts for synthetic data generation

Greetings
Provide a list of 10 phrases commonly used as greetings or well-wishes, where one of the words could also refer to a person's emotion in a different context. Output only the 10 sentences, separated by a newline character, and nothing else.

Social pleasantries
Provide a list of 10 phrases commonly used as social pleasantries, where one of the words could also refer to a person's emotion in a different context. Output only the 10 sentences, separated by a newline character, and nothing else.

Physical vs Emotional
STEP 1: Think of all word derivations of the words ""pain"" and ""hurt"".
STEP 2: Provide a list of 10 text message-style sentences where these words (as a noun, verb, adjective, or adverb) describe emotional distress. Then, provide 10 more sentences where these same words describe physical discomfort or pain in the body. Use casual, text message-style language, and vary the reading level.
STEP 3: Output only the 20 sentences, separated by a newline character and nothing else. Don't output the words you thought of, and don't separate the two sections of 10 sentences.

Subject-experiencer vs. Object experiencer psych verbs
STEP 1: Think of 10 subject-experiencer psych verbs.
STEP 2: Write 10 text message-style sentences where the person who writes the message is the subject of the verb. Then, write 10 text message-style sentences where the person who writes the message is the object of the verb (i.e., someone else is the subject and experiences the emotion toward them). Use casual, text message-style language, and vary the reading level.
STEP 3: Output only the 20 sentences, separated by a newline character and nothing else. Don't output the verbs you thought of, and don't separate the two sections of 10 sentences.

STEP 1: Think of 10 object-experiencer psych verbs.
STEP 2: Write 10 text message-style sentences where the person who writes the message is the object of the verb. Then, write 10 text message-style sentences where the person who writes the message is the subject of the verb (i.e., someone else is the object and that other person experiences the emotion). Use casual, text message-style language, and vary the reading level.
STEP 3: Output only the 20 sentences, separated by a newline character and nothing else. Don't output the verbs you thought of, and don't separate the two sections of 10 sentences.

Psychiatric diagnosis or symptoms vs. emotions
Write 20 text message-style sentences using the words 'depressed' or its derivatives (noun, adjective, adverb) and 'anxious' or its derivatives (noun, adjective, adverb). In the first 10 sentences, these words should clearly describe a psychiatric diagnosis, symptom, or be associated with professional mental health care (e.g., therapy, medication, etc.). In the second 10 sentences, the words should either describe an emotion, or their meaning should be ambiguous, leaving it unclear whether they refer to an emotion or a symptom of a mental health condition. Write the sentences in casual, text message-style language, and vary the reading level. Output only the 20 sentences, separated by a newline character and nothing else.

Meta-emotion vs direct emotion
STEP 1: Think of 5 verbs of preference that express emotions, including positive and negative emotions.
STEP 2: Write 5 text message-style sentences where the verb of preference reflects a meta-emotion, focusing on feelings or emotional experiences rather than external subjects. Then, write another 5 text message-style sentences where the verb of preference directly describes an emotion towards a person or object. Use casual, text message-style language, and vary the reading level.
STEP 3: Output only the 10 sentences, separated by a newline character and nothing else.

Figure S3: Further prompts for synthetic data generation.

Out of sample predictions on comments from Reddit

In the examples, we use the following notation:

- **Words that in other contexts could be emotions but are not detected as such by the model.**
- **Spans detected as emotions by the model.**

---

love
If you can't handle a long book, don't take it away from those of us who **love** them...
I had a bf I **loved** dearly and I wanted kids and he didn't.
I'd **love** to get out more and visit other countries.
I have a good friend who rides the FJ-09, he's in **love** with that bike.

worry
This is really **worrying**, isn't it?
I **worried** about this especially since we had sex daily for the first four months of our relationship and it tapered off after that.
If you can quit your job and not **worry** about it you are rich.
No **worries**!

frustration
I would have saved money, time and **frustration** for sure.
You know what **frustrates** me quite a lot?
I quit the company in 2006 and I don't know what they have been up to since then, but I wouldn't be too surprised if they were similar systems in place elsewhere to identify when a customer is particularly **frustrated**.
Shit, I can remember people breaking equipment out of **frustration** (it happens), but never verbally assaulting an official like that.

pain
As for the tradition, I have three kids and taking them anywhere is a **painful** affair.
it's been 20 years, and i will never get over the **pain** and **regret** of letting his mistakes come between us.
She also noted **pain** that made it hard to walk to next day, which is also associated with the pelvic floor.
Can you imagine what **pain** someone must be in to get to that point?

shock
I was **shocked** how much better my fiver sounded compared to my z3 coupe
Waht a **shocking** story.
That you think it's acceptable to **shock** an 11 year old with 50,000 volts tells me we have very different ideas about policing.
You went into **shock**!

disgust
It's **disgusting** that you refuse to believe a woman who has clearly been attacked.
The amount of people both blindly supporting and blindly hating Eminem in this sub is **disgusting** lmao.
I don't even find these guys **disgusting** like pretty much every other bug.
Worst case scenario, it would be small and hairy, discolored and warty and I'd be absolutely **disgusted**.

Figure S4: Out of sample predictions on comments from Reddit. The comments were selected from the reddit-corpus-small dataset within the ConvoKit toolkit (Chang et al., 2020) for illustrative purposes.

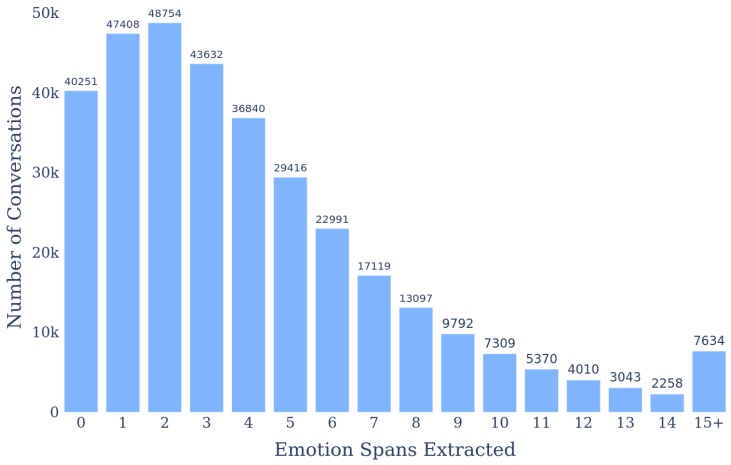
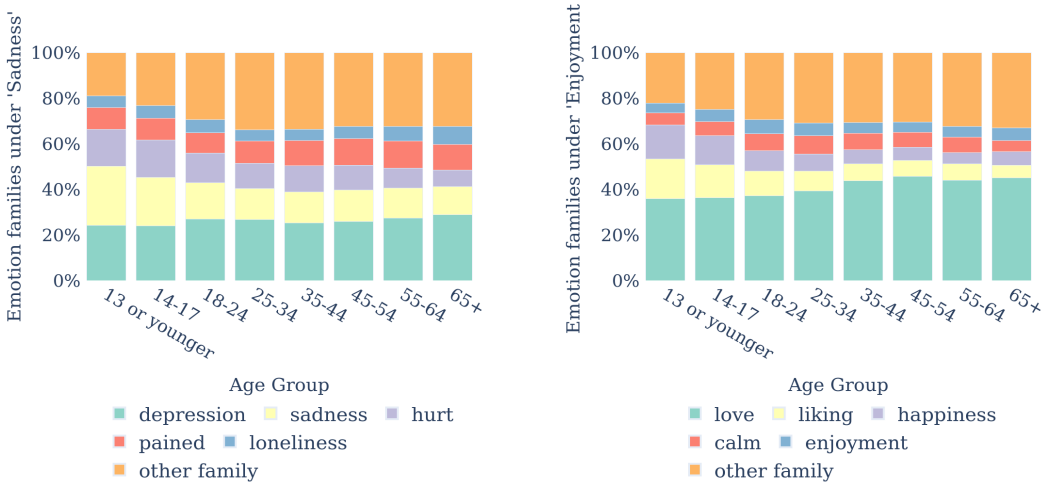
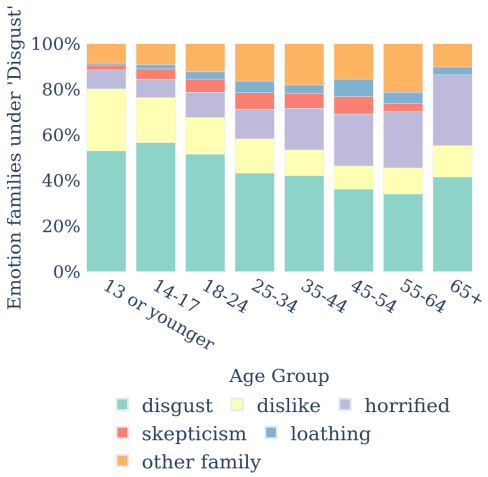


Figure S5: Distribution of the number of extracted emotion spans in a conversation.

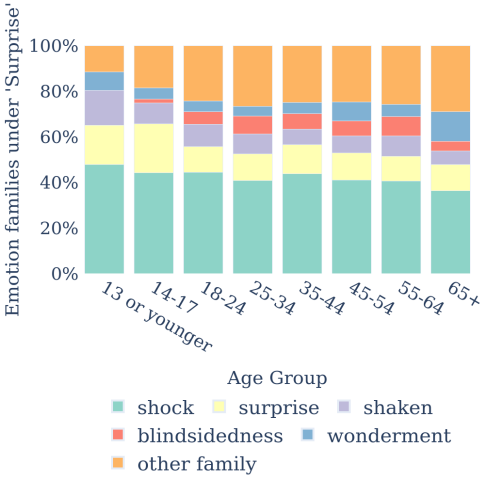


(a) Sadness

(b) Enjoyment



(c) Disgust



(d) Surprise

Figure S6: Relative frequencies of top emotion word families across age groups under supercategories ‘Sadness’,‘Enjoyment’, ‘Disgust’ and ‘Surprise’.