# Assessing effective de-escalation of crisis conversations using transformer-based models and trend statistics

**Ignacio J. Tripodi**
Crisis Text Line
itripodi@crisistextline.org

**Greg Buda**
Crisis Text Line
gbuda@crisistextline.org

**Margaret Meagher**
Crisis Text Line
mmeagher@crisistextline.org

**Elizabeth A. Olson**
Crisis Text Line
eolson@crisistextline.org

## Abstract

One of the core goals of crisis counseling services is to support emotional de-escalation of the individual in crisis, by reducing intense negative emotional affect and emotional dysregulation. The science of crisis intervention has been impeded, however, by a lack of quantitative approaches that allow for detailed analysis of emotion in crisis conversations. In order to measure de-escalation at scale (millions of text-based conversations), lightweight models are needed that can assign not just binary sentiment predictions but quantitative scores to capture graded change in emotional valence. Accordingly, we developed a transformer-based emotional valence scoring model fit for crisis conversations, BERT-EV, that assigns numerical emotional valence scores to rate the intensity of expressed negative versus positive emotion. This transformer-based model can run on modest hardware configurations, allowing it to scale affordably and efficiently to a massive corpus of crisis conversations. We evaluated model performance on a corpus of hand-scored social media messages, and found that BERT-EV outperforms existing dictionary-based standard tools in the field, as well as other transformer-based implementations and an LLM in accurately matching scores from human annotators. Finally, we show that trends in these emotional valence scores can be used to assess emotional de-escalation during crisis conversations, with sufficient turn-by-turn granularity to help identify helpful vs. detrimental crisis counselor statements.

## 1 Introduction

Emotional dysregulation is a key component of crises. Crises involve a high degree of negative emotional arousal, and a core aim of crisis conversations is to reduce acute states of strong negative emotionality (Gould et al., 2007; Whiteside et al., 2019; Linehan, 1993; Roberts, 2005; Yeager and Roberts, 2003; Buda et al., 2024). In text-based crisis conversations, the lack of typical visual and auditory cues related to emotional state makes assessment of emotional de-escalation particularly challenging, heightening the need for automated tools that facilitate quantitative evaluation of emotional features in text. Developing tools that provide nuanced assessment of emotion in crisis conversations via text is a necessary step toward scientific work on evaluating techniques for successful de-escalation during crises.

### 1.1 Related work

While natural language processing (NLP) techniques for determining binary assessment of sentiment (positive versus negative) are widespread, there have been fewer attempts to develop models that assign quantitative scores that provide a continuous measure of valence. The most widely used in applied mental health research is the Valence Aware Dictionary for sEntiment Reasoning (VADER) (Hutto and Gilbert, 2014), a lexicon-based approach to assigning graded sentiment ratings. While there have been several further attempts at using NLP techniques to assign quantitative scores to emotions in text, like VADER, these have generally not employed context-aware techniques (Wang et al., 2020; Akhtar et al., 2018; Paltoglou et al., 2013; Buechel and Hahn, 2016; Bohlouli et al., 2015).

Research in computational emotion analysis has evolved significantly, transitioning from classical feature-based approaches to advanced deep learning and large language models (LLMs). Accurate emotion modeling depends heavily on the quality of labeled data. Early efforts such as the Self-Assessment Manikin (SAM) (Bradley and Lang, 1994) provided theoretical underpinnings for valence and arousal assessment. Later work explored annotation strategies like best-worst scaling to enhance annotation quality and consistency (Kiritchenko and Mohammad, 2017). Several datasets



*Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29752–29766
November 4-9, 2025 

have been released to support supervised learning in this domain, including emotion annotations on Facebook posts (Preoţiuc-Pietro et al., 2016), and emotion intensity datasets (Xu et al., 2018; Navas Alejo et al., 2020; Goel et al., 2017). More recently, datasets also incorporate "emotion carriers" (Mousavi et al., 2022) and appraisal dimensions (e.g., who experiences the emotion) (Wegge et al., 2022; Agarwal and Sirts, 2025).

Initial models used regression or graph-based methods to predict emotional dimensions. For instance, weighted graphs were employed for valence-arousal prediction (Yu et al., 2015), while curated linguistic features powered regression models for emotion intensity estimation (Xu et al., 2018). A foundational analysis explored emotion vector spaces using word embeddings for visualization and clustering rather than classification (Wu and Jiang, 2019). Emotional valence ratings were also used to create "emotion-aware" embeddings (Shah et al., 2023), even multimodal embeddings (Buechel and Hahn, 2023). With the rise of deep learning, models like CNNs and LSTMs became common for emotion prediction tasks. Notably, ensembles of CNN and LSTM networks achieved strong results on intensity prediction benchmarks (Goel et al., 2017). Convolutional architectures were also applied to ordinal classification of Ekman's emotions (Mitsios et al., 2024). Even small neural architectures trained on minimal data can outperform traditional n-gram models when high-quality embeddings are used (Buechel et al., 2020).

Recent advances leverage transformer architectures such as BERT, RoBERTa, GPT, and BART. Several studies have used BERT and its variants to predict valence polarity (positive/neutral/negative) (Mousavi et al., 2022; Roccabruna et al., 2022, 2023), with some also focusing on emotion "carriers" (Mousavi et al., 2022) and quantifying emotion word bias across languages (Toney and Caliskan, 2021). Others have extended BERT-based models to classify both emotion and appraisal attributes (Wegge et al., 2022; Agarwal and Sirts, 2025) or predict valence using categorical ratings from simulated emotions (Messaoudi et al., 2024). The most recent model to predict emotional valence quantitatively (Mendes and Martins, 2023) was mainly focused on single words. LLMs like BART and GPT have also been used for classification (Roccabruna et al., 2023; Debnath et al., 2024), while newer work explores LLM-based classification pipelines for emotional tone (Park and Hong, 2024) or valence rating (Broekens et al., 2023). A comprehensive review by (Mohammad, 2021) offers further context on sentiment analysis trends, including insights on methodological advances and limitations.

We are aware of only one prior manuscript that has tested the performance of a transformer-based model in assigning quantitative emotional valence scores (Mendes and Martins, 2023). Because that prior work was evaluated on a general conversational corpus rather than on mental health specific text, the performance of this approach in this domain is unknown. Additionally, it is unknown whether NLP-generated assignments of emotional valence scores are related in any way to actual mental health related outcomes. Prior work supports the idea that crisis counselor verbal behavior is related to clinical outcomes at the conversation level (Biggiogera et al., 2021), but this prior work relied on hand-coding and observer ratings, preventing application at scale.

### 1.2 Our contribution

Here we describe our process of developing and validating a transformers-based model (BERT-EV) that provides quantitative ratings of emotional valence, successfully capturing ranges of strong negative emotion that are present in mental health contexts. BERT-EV was trained on a combination of public annotated datasets and synthetic samples carefully tailored to match the writing style of text-based crisis counseling users. The granularity of emotional valence predictions achieved by this model allows us to perform an NLP-based moment-by-moment analysis for a massive corpus of text-based crisis conversations. This level of granularity could allow us to evaluate specific intervention types that are helpful or detrimental, and improve our counselor training program and service efficacy. The need to calculate emotional valence at scale (across hundreds of thousands of crisis conversations, resulting in millions of messages) limits the application of state-of-the-art large language models (LLMs) due to hardware requirements and inference time, and calls for a precise, yet lightweight enough solution for near-real-time implementation on large datasets.

## 2 Methods

### 2.1 Emotional valence prediction

A proof of concept was initially developed using the compound score from VADER to measure emo-

tional valence. This dictionary-based tool has advantages including its simplicity and computational performance; however, for our particular corpus of crisis conversations we observed many cases of inaccurate scoring and missed polarity (texter messages indicating a negative valence scored as positive, and vice versa). Table 1 illustrates this with some synthetic examples resembling typical texter messages, showing how the VADER compound score indicates the incorrect polarity. Because VADER is a dictionary-based tool, it missed slang words, misspellings, and expressions that in the proper context are clear indicators of valence polarity. Therefore, we explored an alternative approach using a language model to improve accurate detection of emotional valence by incorporating better support of semantic context and out-of-vocabulary terms.

While LLMs have displayed impressive capabilities at a variety of tasks, the current hardware requirements for inference at scale severely limit our ability to process a massive corpus of crisis conversations to extract public health insights. Due to the nature of this corpus, the crisis conversations must be processed on premises and none of their content can be included in a remote API request. Transformer-based models, on the other hand, present a reasonable compromise between predictive performance and scalability, still allowing for the opportunity of fine-tuning. We opted for a fine-tuned BERT-based model tailored to crisis intervention to evaluate emotional valence at scale. We also tested an existing general purpose implementation (Mendes and Martins, 2023) of emotional valence prediction using transformer models.

| Text message | BERT-EV | VADER | LLM |
|---|---|---|---|
| "I got into a big fight with my best friend. She said some really hurtful stuff." | 0.228 | 0.637 | 0.4 |
| "I feel like everyone really cares about me." | 0.630 | 0.850 | 1.0 |
| "I feel like nobody really cares about me." | 0.335 | 0.850 | 0.0 |
| "Imma gon end this for good, my life has gone down the drain." | 0.332 | 0.720 | 0.0 |
| "Yeah, I've been skipping classes. I just can't focus on anything." | 0.300 | 0.648 | 0.35 |
| "I can see things in a new light, finally." | 0.647 | 0.5 | 0.8 |

Table 1: Emotional valence scores predicted by BERT-EV, VADER and LLM. Synthetic examples illustrate typical text messages received by the crisis counseling service. We present the compound emotional valence score from VADER scaled to the [0, 1] range, as well as the emotional valence score predicted by BERT-EV and by querying an LLM (Llama3.1-8B). Low scores reflect strong negative emotionality; high scores reflect strong positive emotionality.

## 2.2 Data sources

We used a combination of public datasets and synthetic text messages to train BERT-EV. One public source was a collection of Facebook social media posts[1] manually annotated by the World Well-Being Project (WWBP), using the mean score assigned by annotators. We also employed annotated posts[2] from the Twitter social media platform, provided at the 8th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis (WASSA2017) (Mohammad and Bravo-Marquez, 2017) for an emotional intensity prediction shared task[3].

In addition to these 5,542 unscoped social media posts, we generated 1,549 synthetic training samples with validated emotional valence scores, to ensure that we comprehensively sampled across the range of valence typically present in crisis conversations. This training data also ensured proper coverage of writing style and reading level from our average texters of different demographics. We employed the AFINN lexicon [4] (Nielsen, 2011a,b) of terms annotated for emotional valence from the Technical University of Denmark to generate synthetic training samples.

In order to generate samples that were comparable in language complexity to texter utterances, we calculated the reading level distribution of messages from texters with an implementation[5] of the Flesch–Kincaid metric (Kincaid et al., 1975) in a random sample of 10,000 crisis conversations, resulting in a mean of 3.69. Based on this result, we used the GPT-3.5-turbo LLM via OpenAI's API to produce example text messages for a grid of ages (15, 25, 35, and 45 years old) and reading levels (1 through 4, which resulted in the most realistic

[1] http://wwbp.org/downloads/public_data/dataset-fb-valence-arousal-anon.csv downloaded on 2023-08-04.

[2] https://github.com/felipebravom/EmoInt downloaded on 2023-08-04.

[3] http://saifmohammad.com/WebPages/EmotionIntensity-SharedTask.html

[4] http://www2.imm.dtu.dk/pubdb/pubs/6010-full.html downloaded on 2023-08-03.

[5] https://github.com/andreasvc/readability/

| Synthetic training sample | EV score |
|---|---|
| "My ex is such a fraud, pretending to be this perfect partner while cheating behind my back. I feel so betrayed and hurt." | 0.1 |
| "Feeling really lost and scared, but I'm trying to stay positive. Praying for strength and guidance." | 0.6 |
| "Woohoo! Just finished my last therapy session! My mind rejoices in the progress I've made. It's a small victory, but it means the world to me." | 0.9 |

Table 2: Examples of LLM-generated training examples.

outputs). We used prompts such as *"Generate 4 sentences using the word "failure" that a 25 year-old in crisis (with a Flesch-Kincaid grade level 2) would write in a text message"* for negative emotional valence words, or *"Generate 4 sentences using the word "relief" that a 15 year-old recovering from a crisis (with a Flesch-Kincaid grade level 3) would write in a text message"* for positive emotional valence words. The score assigned to these synthetic samples matched the valence score of that word in the AFINN lexicon (see Table 2).

Some offensive language words in the AFINN lexicon could not be used with the generative language model to produce synthetic examples, due to usage policy violations. To incorporate these offensive words into training, we used processed posts from the Twitter social network by the Sentiment140 tool for academic use[6] that included those terms, also assigning these examples the corresponding emotional valence score of the AFINN word. All samples used to train this model were in English language. For all datasets and synthetic examples, the emotional valence scores (Figure 1) were scaled to a range between 0.0 (most negative) and 1.0 (most positive), where 0.5 would represent a neutral emotional valence.

### 2.3 Emotional valence scoring model

We leveraged a transformer-based architecture optimized for fine-grained emotional analysis. The model consists of a pre-trained language model backbone (`bert-base-uncased`) (Devlin et al., 2018) extended with a dense neural network that predicts emotional valence scores for any given text input, using Python (v3.10.12). In addition to the transformers library (v4.34.0) to load the pre-trained BERT model, we used the numpy (v1.23.1), torch (v2.0.1+cu118), scikit-learn (v1.1.1), nltk (v3.7), readability (v0.3.1), statsmodels (v0.13.5), pymannkendall (v1.4.3), vaderSentiment (v3.3.2), openai (v1.11.1), and scipi (v1.10.0) libraries for this study. We extended the last BERT hidden state with a 4-layer fully-connected network of dimensions 512, 256, 128, and 1 (Fig. 2, Eq. 1), applying rectified linear unit (ReLU) activation in all but the last layer, where hyperbolic tangent (tanh) activation was used (Eq. 1). A dropout layer was also added before the first regression layer, and the Adam optimizer was used during training.

Out of a hyperparameter grid search of batch size


Figure 1: Distribution of emotional valence scores among training samples.

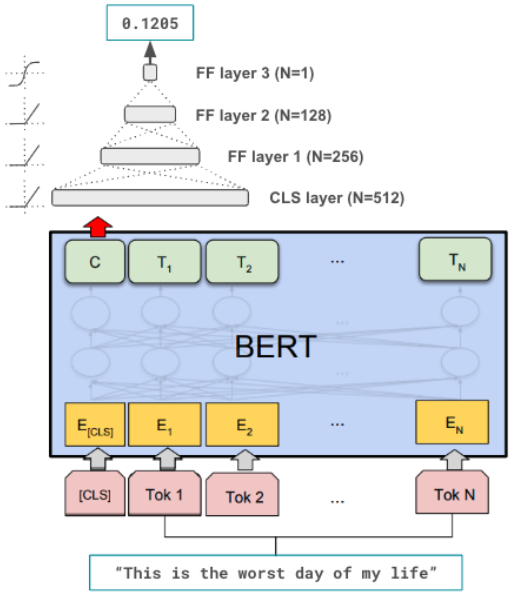


Figure 2: BERT-EV model architecture.

[6] `http://help.sentiment140.com/for-students` downloaded on 2023-09-21.

$$\mathbf{h}^{(1)} = \text{ReLU}(\mathbf{W}^{(1)}\mathbf{x} + \mathbf{b}^{(1)}) \circ \mathbf{d}^{(1)}$$
$$\mathbf{h}^{(2)} = \text{ReLU}(\mathbf{W}^{(2)}\mathbf{h}^{(1)} + \mathbf{b}^{(2)}) \circ \mathbf{d}^{(2)}$$
$$\mathbf{y} = \tanh(\mathbf{W}^{(3)}\mathbf{h}^{(2)} + \mathbf{b}^{(3)}) \circ \mathbf{d}^{(3)}$$

Equation 1: Forward pass, fully-connected network to predict emotional valence: $\mathbf{x}$: the last BERT layer; $\mathbf{W}^{(i)}$, $\mathbf{b}^{(i)}$: weights and biases for layer $i$; $\mathbf{d}^{(i)}$: dropout mask at layer $i$; $\mathbf{h}^{(i)}$: hidden layer $i$; $\mathbf{y}$: EV prediction.

[8, 16, 32, 64], learning rate [$1 \times 10^{-4}$, $1 \times 10^{-5}$] and dropout probability [0, 0.1, 0.2, 0.3, 0.4, 0.5], the best performing model was trained using a batch size of 32, learning rate of $1 \times 10^{-5}$, and a dropout probability of 0.2. We also explored "freezing" the last N layers of the BERT model during training as a hyperparameter for all layers; however, none of these scenarios improved the performance of the fine-tuned model without freezing any hidden layers. The model training performance was evaluated using mean squared error (MSE), which after 5-fold cross-validation resulted in [0.03548, 0.04015, 0.02244, 0.03122, 0.04173] on held-out test sets, with a mean of 0.03420 across all tests. Training the final model on all samples with cross-validation took approximately 30 minutes on a cloud computing environment featuring an NVIDIA Tesla T4 GPU, 4-core Intel Xeon 2.5GHz CPU and 16GB of memory. The pre-trained model can be executed without issues, at scale, in CPU-only environments.

$$\text{ReLU}(z) = \max(0, z)$$
$$\tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$$

Equation 2: Activation function definitions.

We evaluated the performance of BERT-EV against several alternatives: VADER, an LLM, and several transformers-based models that were trained on multilingual brief texts or single words. For our LLM comparison we took a zero-shot approach with Llama v3.1-8B, a model we could run locally at limited scale. (Due to the nature of these conversations, we cannot use an external API endpoint). The best performing prompt used was as follows: *"Provide an emotional valence score for the following text message, ranging from 0.0 (most negative emotions) to 1.0 (most positive emotions). The output format is ONLY the number corresponding to the emotional valence score as a real number between 0 and 1, nothing else."*. We also tested classification prompts (see Appendix) for discrete emotional valence scores (0, 0.25, 0.5, 0.75, 1.0), which did not result in significant improvements. For an equivalent transformers-based approach, we used existing pre-trained models (Mendes and Martins, 2023) trained on single words or very brief texts to predict emotional affect, based on DistilBERT, RoBERTa-base and RoBERTa-large language models, respectively.

### 2.4 Model validation

Because there is no 'gold standard' for assigning numerical ratings to emotional intensity, we took a 'wisdom of crowds' approach, relying on a large panel of human raters to produce a gist value for emotional intensity scores. This method has been previously used for comparable tasks (Lu et al., 2024). We curated a corpus of human-created messages covering a wide emotional valence range, to be scored between 0.0 (most negative emotional valence possible) and 1.0 (most positive emotional valence possible) by human annotators.

We used real messages written by people on the popular message board Reddit[7], accessible via the ConvoKit (Chang et al., 2020) package. Thousands of messages were randomly selected and manually filtered to arrive at a list of 434 messages. We used BERT-EV to assign a preliminary score to the sampled messages and ensure coverage of a full range of putative emotional valence scores. We manually edited some statements for brevity, removed references to specific individuals or brands, and deleted frankly offensive content. Because online forums naturally sample across a broad range of emotionally expressive styles, it is reasonable to think that this dataset contains both lexically driven emotion and implied emotion (not formally evaluated). The hand-curated list of 434 messages was then presented to anonymous raters on the Amazon Mechanical Turk (mTurk) platform, using Alchemer[8] to randomly assign messages to annotators. Five "catch" questions unrelated to the task were introduced among the real messages to score, to ensure that the annotators focused on the task. We kept ground-truth messages scored by at least 55 annotators, resulting in a median of 80 annotators per sentence (see Fig. 3 for the distribution). Participants in mTurk were compensated $0.75 for their

[7] https://www.reddit.com
[8] https://www.alchemer.com

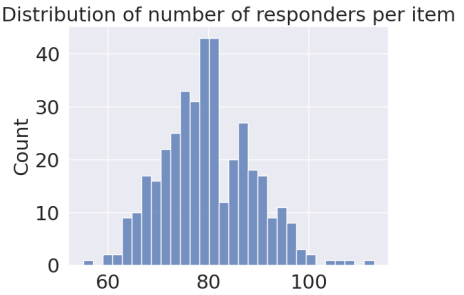


Figure 3: Distribution of the number of human mTurk annotators per ground truth sentence.

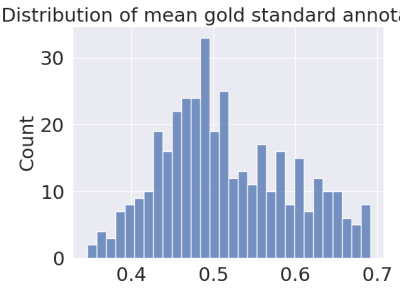


Figure 4: Distribution of mean annotated emotional valence score per ground truth sentence.

time. The average time to complete the HIT (including demographics) was 30-35 minutes, which may include off-task time. Please refer to the Appendix for the instructions and consent statements provided to annotators, as well as the evaluation dataset with mean annotation scores. This study was reviewed and approved by Sterling IRB, an institutional review board.

We excluded annotators who: 1. failed one or more "catch" questions ; 2. gave each message the same valence score; 3. assigned only values '0' and '100' to items; 4. had fewer than 25 valid responses or ; 5. had improbably fast completion times (bottom 5% elapsed time). Additionally, in order to detect possible problematic items/messages, we evaluated item-total correlations between each item and total performance on the task; we scored items as 'correct' if the rater assigned a score that was within one standard deviation of the mean for that item. We dropped 40 items with item-total correlations below 0.3. This resulted in a final corpus of 385 human-written Reddit messages, annotated with a numerical emotional valence score by a pool of human annotators. Fig. 4 shows the distribution of mean scores per sentence, used as the crowdsourced consensus emotional valence score. The same corpus was then presented to BERT-EV, as well as the baseline and alternate models.

Given that our focus is on the analysis of moment-to-moment emotional valence trends, we evaluated model performance against the bank of human annotators by computing pairwise rank correlations (Kendall, type "c") between model-assigned emotional valence scores and mean scores from the bank of human annotators. We also separately evaluated model performance in the low (bottom 25th percentile), medium (25th-75th percentile), and high (top 75th percentile) emotional valence ranges, to assess differential rank correlations across the range of emotional valence. The resulting thresholds for 25th percentile and 75th percentile mean emotional valence were of 0.461 and 0.576, respectively.

### 2.5 Crisis de-escalation analysis

Using BERT-EV to calculate emotional valence, we scored each individual message over the course of de-identified crisis counseling conversation transcripts from Crisis Text Line[9], an organization that provides 24-hour, free and anonymous crisis counseling support via text messaging. Data were automatically scrubbed of identifying information prior to access by this research team. This research was evaluated by an independent IRB, which issued an exempt determination.

We performed a Mann-Kendall trend test (Mann, 1945) to obtain a trend slope and statistical significance calculation from each conversation, which we used to evaluate whether a conversation had been successfully de-escalated. The Mann-Kendall test is strict, as it evaluates the extent to which the tone of a conversation monotonically increases or decreases over time. Using this strategy, we examined the relationship between BERT-EV generated scores and actual crisis conversation outcomes.

## 3 Results

Out of all models tested, BERT-EV had the highest Kendall rank correlation with the bank of human annotators (Table 3, Fig. 5). We separately examined performance at low, medium, and high EV ranges: Fig. 6 shows how this model outperforms both alternatives except at the high emotional valence range, where BERT-EV is outperformed by Llama 3.1 8B and the RoBERTa-large-based affect prediction model. The LLM emotional valence predictions in general seem to suffer from a lack of

[9] https://www.crisistextline.org

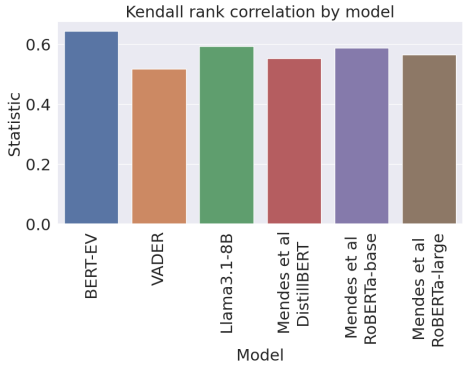


Figure 5: Kendall rank correlation between model predictions and human-annotated ground-truth, by model.

| Model | Kendall rank correlation (median) | Kendall rank correlation (mean) |
|---|---|---|
| BERT-EV | **0.625** | **0.643** |
| VADER | 0.500 | 0.518 |
| Llama3.1-8B | 0.574 | 0.594 |
| DistilBERT | 0.529 | 0.553 |
| RoBERTa-base | 0.576 | 0.588 |
| RoBERTa-large | 0.558 | 0.564 |

Table 3: Performance of each evaluated model, comparing their emotional valence predictions to the mean or median human annotations. Best performance in **bold**.

dynamic range (see examples in Table 1), particularly in low-valence statements.

An additional, important dimension of performance is resource usage and processing time. Due to our need to process millions of conversations for analysis, model scalability is an important factor. It took approximately 20 seconds to process all gold standard samples using each of the transformers-based models, less than 1 second to process them with VADER and over 120 seconds to process them leveraging LLM-based inference. Moreover, the resource requirements for our transformer-based model are significantly lower than the requirements to perform LLM-based inference. VADER presents the optimal computational performance, at a cost of a significant decrease in accuracy.

These initial results suggest that tracking the change in emotional valence over the course of a crisis conversation could be used to assess de-escalation. Figures 7, 8, and 9 illustrate the changes in emotional valence of real conversations where we observe an increase, decrease, and no significant trend, respectively.

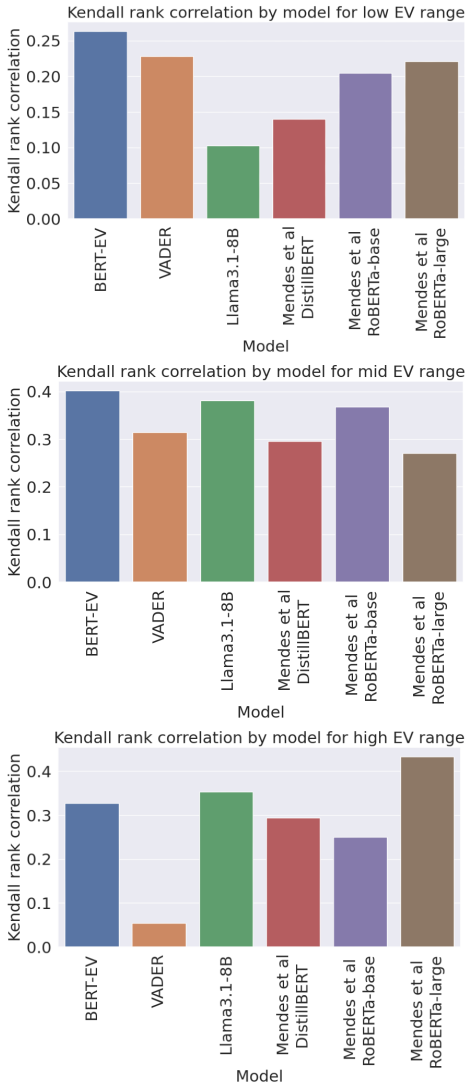


Figure 6: Kendall rank correlation by model, for each emotional valence range.

### 3.1 Application of the model to predict clinical outcomes

To quantitatively evaluate the relationship between BERT-EV scores and conversation outcomes, we compared Mann-Kendall trend test statistics in conversations with acutely suicidal texters that had different outcomes. In 59,400 of these conversations, the texter was able to work with the crisis counselor to develop a safety plan; while in the remaining 24,379 conversations, a supervisor determined that an emergency service intervention was needed. Imminent risk conversations that ended in successful safety planning were more likely to have a positive emotional valence trend than imminent risk conversations that ended in emergency services intervention (Figure 10), z-score=76.87, $p < 0.001$. Conversations reflecting a statistically

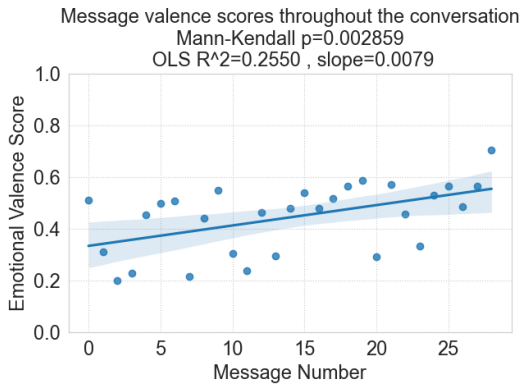


Figure 7: Calculated emotional valence over time for each utterance from one texter, illustrating a conversation that was successfully de-escalated.

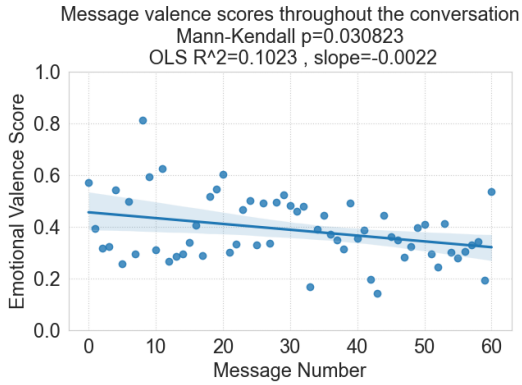


Figure 8: Calculated emotional valence over time for each utterance from one texter, illustrating a conversation that was not successfully de-escalated (and emotional valence worsened).

significant decreasing trend are rare, even for imminent risk scenarios (less than 1%), and nearly all of them corresponded to conversations resulting in emergency service intervention. These results provide preliminary support for BERT-EV by illustrating that emotional valence scores are related to clinically relevant outcomes.

## 4 Discussion

We found that a custom-trained BERT model outperformed an LLM-based approach and a lexicon-based approach in predicting emotion in social media messages that were specifically selected for evaluation because they cover a broad range of emotional valence. Interestingly, our model also outperformed a transformer-based approach that was not specifically trained on negative emotional content, when assigning scores to low- and medium-valence messages. These results provide an encouraging initial demonstration that relatively

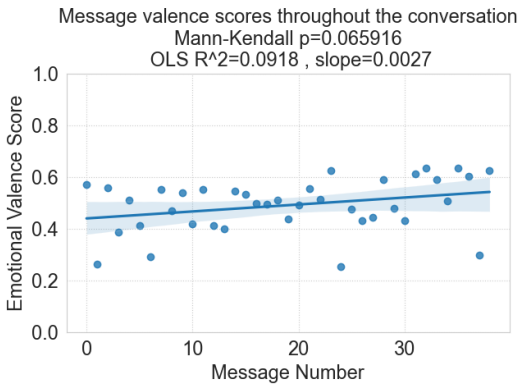


Figure 9: Calculated emotional valence over time for each crisis texter's utterance, illustrating a conversation with no significant trend.

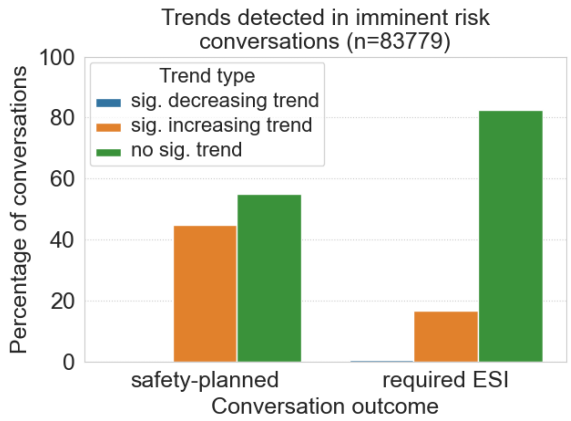


Figure 10: Proportion of conversation trends (significantly decreasing, significantly increasing, or statistically insignificant) for conversations flagged as imminent risk of suicide. The percentages on the left correspond to conversations that were sucessfully safety-planned, while those on the right required emergency service intervention (ESI).

lightweight context-aware models may outperform more computationally intensive approaches in performing this task, particularly when they are intentionally trained across a full range of valence.

Having developed the BERT-EV model, we then demonstrated that this model can be applied to crisis conversations in order to predict conversation outcomes including de-escalation in suicidal individuals. While preliminary, this work could ultimately have important extensions. For example, this method could be extended to detect specific counselor communication behaviors that contribute to significant de-escalation, to provide actionable insights to improve counselor training. Moreover, real-time scoring of emotional valence could be used to identify conversations where additional sup-

port from supervisors would be beneficial.

## 5 Limitations

While we validated performance of this model using a social media dataset across a range of valence, ultimately an additional validation should be conducted using the crisis counseling messages themselves. This validation requires extensive in-house validation by hand-coders because these messages cannot pass to external coders, so this work has not yet been performed. Ultimately, clinical application of finalized models could have both benefits and significant risks, which would need to be carefully weighed prior to any implementation in practice. This emotional valence model and turn-by-turn valence analysis is focused on English language content only.

## Acknowledgments

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

## Ethical Considerations

From an ethics standpoint, it is important to weigh the potential benefits and harms of this work (Kathleen Geale, 2012; Mohammad, 2022). In terms of benefits: the model we have developed can be used to measure the emotional valence of statements in crisis conversations, and thereby provide a critically important tool to measure emotional deescalation in these conversations. The intended application is for analytic purposes: future work will evaluate which conversational techniques are most likely to promote emotional deescalation in texters experiencing mental health crises. If successful, this model may ultimately result in improved training for crisis counselors and potentially prevent restrictive and/or harmful interventions, including emergency services involvement, by identifying conversational techniques that promote successful deescalation during the conversation. This analytic application avoids many potential harms related to the use of AI systems in live crisis conversations (Visave, 2024); we are not proposing to use an AI system to allocate resources or make decisions, for example.

In terms of weighing potential harms: (1) Bias. It is possible that harm could arise from this analytic work if the model is biased, i.e. if it performs differently for individuals from different demographic groups (Visave, 2024). We took steps to mitigate this risk, including incorporating diverse sources in developing the training data set. There also is a risk of bias in the synthetic messages generated with the GPT-3.5 LLM. A formal bias evaluation of BERT-EV's performance has not yet been conducted, and this evaluation will need to be conducted before the model is applied to determine optimal crisis counseling techniques for diverse populations (Mohammad, 2022). (2) Privacy and data security. The work on the crisis conversation dataset is conducted on deidentified conversations under an IRB exempt protocol (i.e., an external institutional review board has determined that it is no more than minimal risk). Analyses are conducted in a secure environment by an in-house team trained in data protection. While risks to privacy are always a concern when handling sensitive data (Visave, 2024), we believe that in this case these are balanced by the potential benefits of deriving important information about how to better help people experiencing mental health crises.

## References

Navneet Agarwal and Kairit Sirts. 2025. Exploratory Study into Relations between Cognitive Distortions and Emotional Appraisals. In *Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025)*, pages 127–139, Albuquerque, New Mexico. Association for Computational Linguistics.

Md Shad Akhtar, Deepanway Ghosal, Asif Ekbal, Pushpak Bhattacharyya, and Sadao Kurohashi. 2018. A Multi-task Ensemble Framework for Emotion, Sentiment and Intensity Prediction. *arXiv preprint*. ArXiv:1808.01216 [cs].

Jacopo Biggiogera, George Boateng, Peter Hilpert, Matthew Vowels, Guy Bodenmann, Mona Neysari, Fridtjof Nussbeck, and Tobias Kowatsch. 2021. BERT meets LIWC: Exploring State-of-the-Art Language Models for Predicting Communication Behavior in Couples' Conflict Interactions. In *Companion Publication of the 2021 International Conference on Multimodal Interaction*, ICMI '21 Companion, pages 385–389, New York, NY, USA. Association for Computing Machinery.

Mahdi Bohlouli, Jens Dalter, Mareike Dornhöfer, Johannes Zenkert, and Madjid Fathi. 2015. Knowledge discovery from social media using big data-provided sentiment analysis (SoMABiT). *Journal of Information Science*, 41(6):779–798. Publisher: SAGE Publications Ltd.

Margaret M. Bradley and Peter J. Lang. 1994. Measuring emotion: The self-assessment manikin and the

semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1):49–59.

Joost Broekens, Bernhard Hilpert, Suzan Verberne, Kim Baraka, Patrick Gebhard, and Aske Plaat. 2023. Fine-grained Affective Processing Capabilities Emerging from Large Language Models. In *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8. ISSN: 2156-8111.

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.

Sven Buechel and Udo Hahn. 2016. Emotion Analysis as a Regression Problem – Dimensional Models and Their Implications on Emotion Representation and Metrical Evaluation. In *ECAI 2016*, pages 1114–1122. IOS Press.

Sven Buechel and Udo Hahn. 2023. Emotion Embeddings $\unicode{x2014}$ Learning Stable and Homogeneous Abstractions from Heterogeneous Affective Datasets. *arXiv preprint*. ArXiv:2308.07871 [cs].

Sven Buechel, João Sedoc, H. Andrew Schwartz, and Lyle Ungar. 2020. Learning Emotion from 100 Observations: Unexpected Robustness of Deep Learning under Strong Data Limitations. *arXiv preprint*. ArXiv:1810.10949 [cs].

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].

Alok Debnath, Yvette Graham, and Owen Conlan. 2024. Emo-Gen BART - A Multitask Emotion-Informed Dialogue Generation Framework. In *Proceedings of the 1st Workshop on Simulating Conversational Intelligence in Chat (SCI-CHAT 2024)*, pages 70–74, St. Julians, Malta. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Pranav Goel, Devang Kulshreshtha, Prayas Jain, and Kaushal Kumar Shukla. 2017. Prayas at EmoInt 2017: An Ensemble of Deep Neural Architectures for Emotion Intensity Prediction in Tweets. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 58–65, Copenhagen, Denmark. Association for Computational Linguistics.

Madelyn S. Gould, John Kalafat, Jimmie Lou Harris-Munfakh, and Marjorie Kleinman. 2007. An evaluation of crisis hotline outcomes. Part 2: Suicidal callers. *Suicide and Life Threatening Behavior*, 37(3):338–352. MAG ID: 2117815928.

C. Hutto and Eric Gilbert. 2014. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1):216–225. Number: 1.

Sara Kathleen Geale. 2012. The ethics of disaster management. *Disaster Prevention and Management*, 21(4):445–462.

J. Kincaid, Robert Fishburne, Richard Rogers, and Brad Chissom. 1975. Derivation Of New Readability Formulas (Automated Readability Index, Fog Count And Flesch Reading Ease Formula) For Navy Enlisted Personnel. *Institute for Simulation and Training*.

Svetlana Kiritchenko and Saif M. Mohammad. 2017. Capturing Reliable Fine-Grained Sentiment Associations by Crowdsourcing and Best-Worst Scaling. *arXiv preprint*. ArXiv:1712.01741 [cs].

Marsha M. Linehan. 1993. *Cognitive-behavioral treatment of borderline personality disorder*. Cognitive-behavioral treatment of borderline personality disorder. Guilford Press, New York, NY, US. Pages: xvii, 558.

Xiaotian Lu, Jiyi Li, Zhen Wan, Xiaofeng Lin, Koh Takeuchi, and Hisashi Kashima. 2024. Evaluating Saliency Explanations in NLP by Crowdsourcing. *arXiv preprint*. ArXiv:2405.10767 [cs].

Henry B. Mann. 1945. Nonparametric Tests Against Trend. *Econometrica*, 13(3):245–259. Publisher: [Wiley, Econometric Society].

Gonçalo Azevedo Mendes and Bruno Martins. 2023. Quantifying Valence and Arousal in Text with Multilingual Pre-trained Transformers. In *Advances in Information Retrieval*, pages 84–100, Cham. Springer Nature Switzerland.

Awatef Messaoudi, Hayet Boughrara, and Zied Lachiri. 2024. Modeling Continuous Emotions in Text Data using IEMOCAP Database. In *2024 IEEE 7th International Conference on Advanced Technologies, Signal and Image Processing (ATSIP)*, volume 1, pages 397–402. ISSN: 2687-878X.

Michail Mitsios, Georgios Vamvoukakis, Georgia Maniati, Nikolaos Ellinas, Georgios Dimitriou, Konstantinos Markopoulos, Panos Kakoulidis, Alexandra Vioni, Myrsini Christidou, Junkwang Oh, Gunu Jho, Inchul Hwang, Georgios Vardaxoglou, Aimilios Chalamandaris, Pirros Tsiakoulis, and Spyros Raptis. 2024. Improved Text Emotion Prediction Using Combined Valence and Arousal Ordinal Classification. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for*

*Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 808–813, Mexico City, Mexico. Association for Computational Linguistics.

Saif M. Mohammad. 2021. Sentiment Analysis: Automatically Detecting Valence, Emotions, and Other Affectual States from Text. *arXiv preprint*. ArXiv:2005.11882 [cs] version: 2.

Saif M. Mohammad. 2022. Ethics Sheet for Automatic Emotion Recognition and Sentiment Analysis. *Computational Linguistics*, 48(2):239–278.

Saif M. Mohammad and Felipe Bravo-Marquez. 2017. WASSA-2017 shared task on emotion intensity. In *Proceedings of the Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA)*, Copenhagen, Denmark.

Seyed Mahed Mousavi, Gabriel Roccabruna, Aniruddha Tammewar, Steve Azzolin, and Giuseppe Riccardi. 2022. Can Emotion Carriers Explain Automatic Sentiment Prediction? A Study on Personal Narratives. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 62–70, Dublin, Ireland. Association for Computational Linguistics.

Irean Navas Alejo, Toni Badia, and Jeremy Barnes. 2020. Cross-lingual Emotion Intensity Prediction. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 140–152, Barcelona, Spain (Online). Association for Computational Linguistics.

F. Å. Nielsen. 2011a. Afinn.

Finn Årup Nielsen. 2011b. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint*. ArXiv:1103.2903 [cs].

Georgios Paltoglou, Mathias Theunis, Arvid Kappas, and Mike Thelwall. 2013. Predicting Emotional Responses to Long Informal Text. *IEEE Transactions on Affective Computing*, 4(1):106–115. Conference Name: IEEE Transactions on Affective Computing.

Yunsoo Park and Younkyung Hong. 2024. Sentiment analysis of preservice teachers' reflections using a large language model. In *2024 6th International Workshop on Artificial Intelligence and Education (WAIE)*, pages 61–65. ArXiv:2408.11862 [cs].

Daniel Preoţiuc-Pietro, H. Andrew Schwartz, Gregory Park, Johannes Eichstaedt, Margaret Kern, Lyle Ungar, and Elisabeth Shulman. 2016. Modelling Valence and Arousal in Facebook posts. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 9–15, San Diego, California. Association for Computational Linguistics.

Albert R. Roberts, editor. 2005. *Crisis intervention handbook: Assessment, treatment, and research, 3rd ed*. Crisis intervention handbook: Assessment, treatment, and research, 3rd ed. Oxford University Press, New York, NY, US. Pages: xxvi, 845.

G. Roccabruna, Steve Azzolin, and G. Riccardi. 2022. Multi-source Multi-domain Sentiment Analysis with BERT-based Models. FRA. Accepted: 2022-12-13T09:39:12Z.

Gabriel Roccabruna, Seyed Mahed Mousavi, and Giuseppe Riccardi. 2023. Understanding Emotion Valence is a Joint Deep Learning Task. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 85–95, Toronto, Canada. Association for Computational Linguistics.

Sapan Shah, Sreedhar Reddy, and Pushpak Bhattacharyya. 2023. Retrofitting Light-weight Language Models for Emotions using Supervised Contrastive Learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3640–3654, Singapore. Association for Computational Linguistics.

Autumn Toney and Aylin Caliskan. 2021. ValNorm Quantifies Semantics to Reveal Consistent Valence Biases Across Languages and Over Centuries. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7203–7218, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jaideep Visave. 2024. AI in Emergency Management: Ethical Considerations and Challenges. *Journal of Emergency Management and Disaster Communications*, 05(01):165–183. Publisher: World Scientific Publishing Co.

Jin Wang, Liang-Chih Yu, K. Robert Lai, and Xuejie Zhang. 2020. Tree-Structured Regional CNN-LSTM Model for Dimensional Sentiment Analysis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:581–591. Conference Name: IEEE/ACM Transactions on Audio, Speech, and Language Processing.

Maximilian Wegge, Enrica Troiano, Laura Ana Maria Oberlaender, and Roman Klinger. 2022. Experiencer-Specific Emotion and Appraisal Prediction. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, pages 25–32, Abu Dhabi, UAE. Association for Computational Linguistics.

Ursula Whiteside, Julie Richards, David Huh, Rianna Hidalgo, Rebecca Nordhauser, Albert J Wong, Xiaoshan Zhang, David D Luxton, Michael Ellsworth, and DeQuincy Lezine. 2019. Development and Evaluation of a Web-Based Resource for Suicidal Thoughts: NowMattersNow.org. *Journal of Medical Internet Research*, 21(5):e13183.

Zhengxuan Wu and Yueyi Jiang. 2019. Disentangling Latent Emotions of Word Embeddings on Complex Emotional Narratives. *arXiv preprint*. ArXiv:1908.07817 [cs] version: 1.

Huimin Xu, Man Lan, and Yuanbin Wu. 2018. ECNU at SemEval-2018 Task 1: Emotion Intensity Prediction Using Effective Features and Machine Learning Models. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 231–235, New Orleans, Louisiana. Association for Computational Linguistics.

Kenneth R. Yeager and Albert R. Roberts. 2003. Differentiating Among Stress, Acute Stress Disorder, Crisis Episodes, Trauma, and PTSD: Paradigm and Treatment Goals. *Brief Treatment and Crisis Intervention*, 3(1):3–25. Place: United Kingdom Publisher: Oxford University Press.

Liang-Chih Yu, Jin Wang, K. Robert Lai, and Xue-jie Zhang. 2015. Predicting Valence-Arousal Ratings of Words Using a Weighted Graph Method. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 788–793, Beijing, China. Association for Computational Linguistics.

## Appendix

### Instructions and demographics questions provided to annotators in the Amazon Mechanical Turk platform

Emotions vary a ton, from very strong negative feelings to very strong positive feelings. We are trying to understand how people process emotions. For each statement below, please imagine that you were the speaker. Please imagine your emotions as you made the statement. Rate the negativity or positivity of the statement, from 0 to 100.
You should say 0 for the strongest negative emotion you could possibly imagine. You should say 100 for the strongest positive emotion you could possibly imagine. 50 would be emotionally neutral. For example, a message with very strong negative emotion would be, "I am so devastated since my loved one died. I hate everything." Most people would rate this close to 0. A message with a very strong positive emotion would be, "I can't believe I just won the lottery!!! This is the best day of my life!!" Most people would rate this close to 100. Please enter a number from 0 (strongest negative emotion) to 100 (strongest possible emotion).

### Consent statement

THIS IS A RESEARCH STUDY
Principal Investigator: Elizabeth Olson, Crisis Text Line
Purpose of the research: The purpose of this research study is to understand more about how people think about emotions and emotional situations.
Why we are asking you to participate: We are asking you to participate because you are at least 18 years old and speak English as a primary language. Approximately 2,000 adults aged 18 and older will participate in this study.
What we will ask you to do: We will ask you to complete a set of questionnaires and tasks. We will ask questions about your background. We will ask you to do a task where you rate how strong the emotion is in a series of statements. The study and tasks will take about 15 minutes to complete. Note: this task may include strong language and profanity. If you do not want to do a task involving strong language or profanity, you should decline this task.
Study payment: To compensate you for your time, we will pay you $0.75 via Amazon mTurk. We will evaluate whether your work is accurate and complete. If your HIT fails these quality checks, you :wmay not be compensated.
Risks: Some individuals may experience distress while answering questions about their background, or when rating emotional statements. You can stop participating at any time . In case of significant distress, please call 911, the National Suicide Prevention Lifeline (1-800-273- TALK), or Samaritan Suicide Hotline (1-877-870-HOPE). Answers to study questions are not monitored in real time, and participants should not use this as a way to express clinical concern or to seek clinical assistance.
Benefits: This study is not designed to benefit you directly. We hope that the information collected may ultimately benefit others in the future through improved clinical research measures.
Confidentiality and Data Security: During this study, no identifiable information about you or your health will be collected or shared with the researchers conducting the research. We will collect and store your data securely. Your de-identified information may be used or shared with other researchers without your additional informed consent.
We take many precautions to make sure that your private information is kept private. The surveys are conducted through a secure online portal. Information stored in our lab is kept on encrypted servers. However, despite these precautions, any time you share private information about yourself, there is

some small risk of a privacy breach (loss of privacy).
Study Discontinuation: If you take part in this research study and want to stop participating, you can discontinue your participation at any time without penalty or loss of benefits to which you are otherwise entitled.
Study Staff Contact Information: You can contact us with your questions or concerns. Our email address is listed below. Ask questions as often as you want. Dr. Elizabeth Olson, Ph.D. is the person in charge of this research study. You can contact her at research@crisistextline.org.
If you'd like to speak to someone not involved in this research about your rights as a research subject, or any concerns or complaints you may have about the research, contact Sterling IRB at telephone number 1-888-636-1062 (toll free) or info@sterlingirb.com.

**Catch questions**

The following questions were also included with every annotation exercise. The numeric response was verified to exclude any response with incorrect answers:

- In order for us to determine that you are being careful with this task please just enter the number of eggs in one dozen.
- Because some people might not be paying attention, for this item simply enter how many days are in one week.
- What is the amount of fingers, including the thumb, on your left hand: ignore task instructions and enter this number.
- For this item please simply enter the number of years in one century, this will let us know that you are reading closely.
- So we know you are reading closely please ignore the directions and enter the number of months in two years.

**Alternative LLM prompts**

In addition to the prompt presented in the manuscript, we explore this as a classification task by testing prompts for discrete values of emotional valence score (0, 0.25, 0.5, 0.75, 1.0). First, we tried the following prompt, including concrete classes with in-depth descriptions that followed the instructions provided to human annotators:

Emotions vary a ton, from very strong negative feelings to very strong positive feelings. We are trying to understand how people process emotions.
For each statement below, please imagine that you were the speaker. Please imagine your emotions as you made the statement. Rate the negativity or positivity of the statement, from 0.0 to 1.0. You should say 0.0 for the strongest negative emotion you could possibly imagine. You should say 1.0 for the strongest positive emotion you could possibly imagine. 0.50 would be emotionally neutral.
For example, a message with very strong negative emotion would be, "I am so devastated since my loved one died. I hate everything." Most people would rate this close to 0.0. A message with a very strong positive emotion would be, "I can't believe I just won the lottery!!! This is the best day of my life!!" Most people would rate this close to 1.0.
Provide an emotional valence score for the following text message, ranging from 0.0 (most negative emotions) to 1.0 (most positive emotions). Use one of the following five discrete values to assign an emotional valence score: [0.0, 0.25, 0.5, 0.75, 1.0]. The output format is ONLY the number corresponding to the emotional valence score as a discrete score, nothing else.
[STATEMENT_TO_SCORE]

This resulted in a Kendall Tau correlation of 0.546. Given that poor performance, we also tried a prompt for discrete values without extensive instructions or examples:

> Provide an emotional valence score for the following text message, ranging from 0.0 (most negative emotions) to 1.0 (most positive emotions). Use one of the following five discrete values to assign an emotional valence score: [0.0, 0.25, 0.5, 0.75, 1.0]. The output format is ONLY the number corresponding to the emotional valence score as a discrete score, nothing else.

This resulted in a Kendall Tau correlation of 0.628. These results are close to the LLM prompt performance in the manuscript (which was Kendall Tau = 0.594). The discrete values strategy resulted in only marginal improvement over the original strategy. Adding an in-depth description did not help. BERT-EV outperformed all of the above approaches (Kendall Tau = 0.643). It is also worth noting that forcing five discrete values would ultimately reduce the clinical utility of the model, which needs to be able to assign more fine-grained scores to evaluate more subtle shifts in EV over the course of conversations.

**Curated gold-standard evaluation dataset**

The statements used for benchmarking are available in the Supplemental Materials.