"That's so cute!": The CARE Dataset for Affective Response Detection

Jane Dwivedi-Yu
Meta AI
janeyu@meta.com

Alon Y. Halevy
Meta AI
ayh@meta.com

Abstract

Social media plays an increasing role in our communication with friends and family, and in our consumption of entertainment and information. Hence, to design effective ranking functions for posts on social media, it would be useful to predict the affective responses of a post (e.g., whether it is likely to elicit feelings of entertainment, inspiration, or anger). Similar to work on emotion detection (which focuses on the affect of the publisher of the post), the traditional approach to recognizing affective response would involve an expensive investment in human annotation of training data.

We create and publicly release $CARE_{db}$, a dataset of 230k social media post annotations according to seven affective responses using the Common Affective Response Expression (CARE) method. The CARE method is a means of leveraging the signal that is present in comments that are posted in response to a post, providing high-precision evidence about the affective response to the post without human annotation. Unlike human annotation, the annotation process we describe here can be iterated upon to expand the coverage of the method, particularly for new affective responses. We present experiments that demonstrate that the CARE annotations compare favorably with crowdsourced annotations. Finally, we use $CARE_{db}$ to train competitive BERT-based models for predicting affective response as well as emotion detection, demonstrating the utility of the dataset for related tasks.
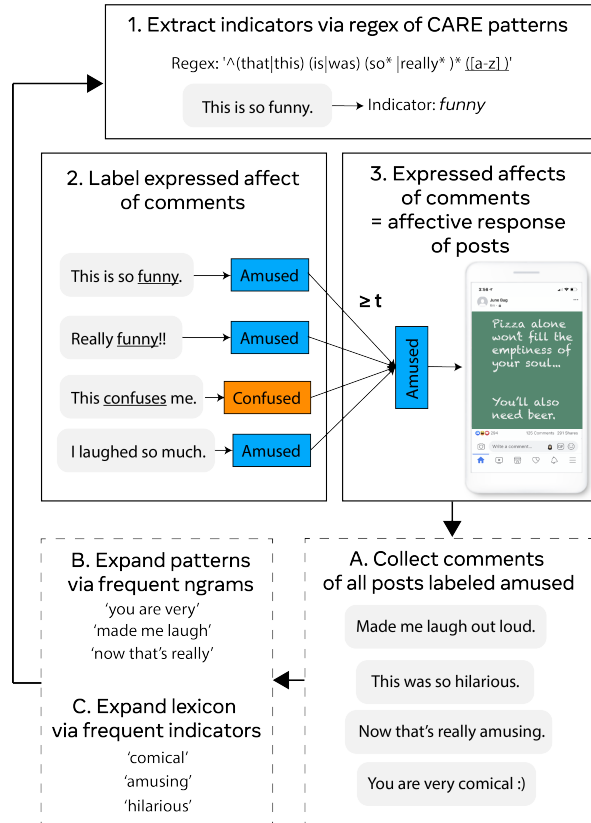
Figure 1: Overview of the CARE Method (pseudo-code in Appendix, Algorithm 1). The top half of the figure (steps 1–3) shows how the affective response to a post is computed by aggregating the expressed affects in comments from users viewing the post. The bottom half of the figure (steps A–C) shows how we expand the collection of CARE patterns and the lexicon based on labels that have been obtained from prior iterations.

1 Introduction

Social media and other online media platforms have become a common means of not only interacting and connecting with others, but also finding interesting, informing, and entertaining content. Users of those platforms depend on the ranking systems of the recommendation systems to show them information they will be most interested in and to safeguard them against unfavorable experiences.

Towards this end, a key technical problem is to predict the *affective response* that a user may have when they see a post. Some affective responses can be described by emotions (e.g., angry, joyful), and others may be described more as experiences (e.g., entertained, inspired). Predicting affective response differs from emotion detection in that the latter focuses on the emotions expressed by the publisher of the post (referred to as the *publisher*

affect in Chen et al. (2014)) and not on the viewer of the content. While the publisher's emotion may be relevant to the affective response, it only provides a partial signal (Dwivedi-Yu et al., 2022), and the two are not always equivalent (see Figure 2 for an illustrative example). Affective response for recommender systems has shown to be critical in several applications such as music, emotional health monitoring systems, product and travel recommendations (Rosa et al., 2015, 2018; Akram et al., 2020; Artemenko et al., 2020; Dwivedi-Yu et al., 2022).
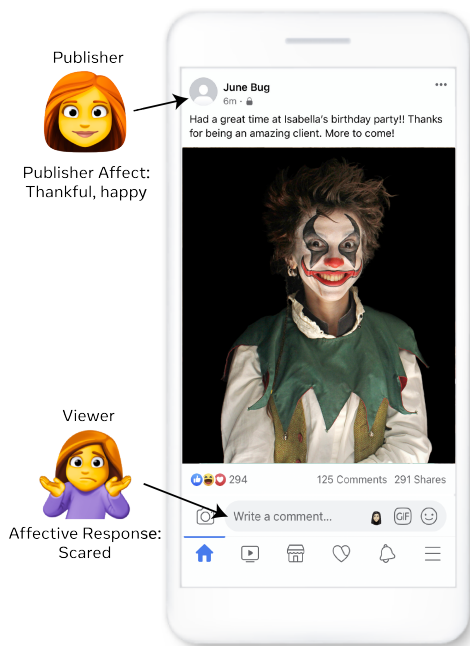

Figure 2: An example case of differing publisher affect and affective response. This work focuses on affective response through signals such as comments and reactions. Post image sourced from Shutterstock (Tapia).

Current approaches to predicting affective response require obtaining training data from human annotators who try to classify content into classes of a given taxonomy. However, obtaining enough training data can be expensive, and moreover, due to the subjective nature of the problem, achieving consensus among annotators can be challenging. Some methods explore inferring responses from physiological data or facial expressions from users, but this is a highly invasive process and can be difficult to scale to multiple users. (Tkalčič et al., 2017, 2019; Angelastro et al., 2019).

This paper introduces the Common Affective Response Expression method (CARE for short), a means of obtaining labels for affective response in an unsupervised way from the comments written in response to online posts. CARE uses patterns and a keyword-affect mapping to identify expressions in comments that provide high-precision evidence about the affective response of the readers to the post. For example, the expression "What a hilarious story" may indicate that a post is humorous and "This is so cute" may indicate that a post is adorable. We seed the system with a small number of high-precision patterns and mappings. We then iteratively expand on the initial set by considering frequent patterns and keywords in unlabeled comments on posts labeled by the previous iteration.

Using CARE, we create the largest dataset to date for affective response, $CARE_{db}$, which contains 230k posts annotated according to 7 affective responses. We validate the effectiveness of CARE by comparing the CARE annotations with crowdsourced annotations. Our experiments show that there is a high degree of agreement between the annotators and the labels proposed by CARE (e.g., in 90% of the cases, at least two out of three annotators agree with all the CARE labels). Furthermore, we show that the CARE patterns/lexicon have greater accuracy than applying SOTA emotion recognition techniques to the comments. Using $CARE_{db}$, we train CARE-BERT[1], a BERT-based model that can predict affective response *without* relying on comments. CARE-BERT provides strong baseline performance for the task of predicting affective response, on par with the SOTA models for emotion recognition. Furthermore, we show that CARE-BERT can be used for transfer learning to a different emotion-recognition task, achieving similar performance to Demszky et al. (2020), which relied on manually-labeled training data.

In summary, our contributions are as follows:

- CARE, a novel method for leveraging the signal present in comments in order to label the affective response of a post, without the need for human annotation.
- $CARE_{db}$, a dataset of 230k annotated posts according to 7 affective responses using CARE.
- Error analysis using human annotations for a sampled set of posts from $CARE_{db}$.
- Quantitative results that demonstrate CARE performs better than a method leveraging a state-of-the-art publisher-affect classifier.

---

[1]The CARE patterns, lexicon, $CARE_{db}$, and CARE-BERT are made available at https://github.com/facebookresearch/care.

- CARE-BERT: A model for labeling affective response from the post text, without the need for comments.
- Transfer learning experiments that demonstrate transferability to different emotion-recognition tasks under low-resource settings.

2 Related work

We first situate our work with respect to previous research on related tasks.

2.1 Emotion detection in text

Approaches to emotion detection can be broadly categorized into three groups: lexicon-based, machine learning, and combinations of the first two. The lexicon-based approach typically leverages lexical resources such as lexicons and encoded rules to guide emotion prediction (Tao, 2004; Ma et al., 2005; Asghar et al., 2017). Though these methods can be fast and interpretable, they are often not as robust and flexible because of the constraints of the lexicon (Alswaidan and Menai, 2020; Acheampong et al., 2020). Additionally, the scope of emotions predicted by these works is usually fairly small, ranging from two to five, and most datasets utilized are usually smaller than 10k, making it unclear if they extrapolate well. Among the ML approaches, many SOTA works employ deep learning methods (Demszky et al., 2020; Felbo et al., 2017; Barbieri et al., 2018; Huang et al., 2019a; Baziotis et al., 2017; Huang et al., 2019b), but while these show significant improvement over prior techniques, they are highly uninterpretable and often require prohibitively large human-labeled datasets to train. In both the lexicon-based approach and the ML-approach, the classes of emotions predicted in these works are usually non-extendable or require additional labeled data.

While there are some commonalities between works in emotion detection and affective response detection, the problems are distinct enough that we cannot simply apply emotion recognition techniques to our setting. Emotion recognition focuses on the publisher affect (the affect of the person writing the text). The publisher affect may provide a signal about the affective response of the reader, but there is no simple mapping from one to the other. For example, being 'angered' is an affective response that does not only result from reading an angry post—it can result from a multitude of different publisher affects (e.g. excited, angry, sympathetic, embarrassed, or arrogant). For some affective responses, such as feeling 'grateful' or 'connected' to a community, the corresponding publisher affect is highly unclear.

2.2 Affective response detection

There have been some works that address affective response through natural language in limited settings, such as understanding reader responses to online news (Katz et al., 2007; Strapparava and Mihalcea, 2007; Lin et al., 2008; Lei et al., 2014). In contrast, our goal is to address the breadth of content on social media. There are works which use Facebook reactions as a proxy for affective response, but these are constrained by the pre-defined set of reactions (Clos et al., 2017; Raad et al., 2018; Pool and Nissim, 2016; Graziani et al., 2019; Krebs et al., 2017). The work described in Rao et al. (2014) and Bao et al. (2012) attempts to associate emotions with *topics*, but a single topic can have a large variety of affective responses when seen on social media, and therefore their model does not apply to our case. Some works in the computer vision community study affective response to images (Chen et al., 2014; Jou et al., 2014); as they note, most of the work in the vision community also focuses on publisher affect.

2.3 Methods for unsupervised labeling

A major bottleneck in developing models for emotion and affective response detection is the need for large amounts of training data. As an alternative to manually-labeled data, many works utilize metadata such as hashtags, emoticons, and Facebook reactions as pseudo-labels (Wang et al., 2012; Suttles and Ide, 2013; Hasan et al., 2014; Mohammad and Kiritchenko, 2015). However, these can be highly noisy and limited in scope. For example, there exist only seven Facebook reactions, and they do not necessarily correspond to distinct affective responses. Additionally, for abstract concepts like emotions, hashtagged content may only capture a superficial interpretation of the concept. For example, searching #inspiring on Instagram will return many photos featuring selfies or obviously inspirational quotes, which do not sufficiently represent inspiration. The work we present here extracts labels from free-form text in comments rather than metadata. The work done in Sintsova and Pu (2016) is similar to our work in that it pseudo-labels tweets and extends its lexicon, but the classifier itself is a keyword, rule-based approach and is heavily re-

liant on the capacity of these lexicons. In contrast, our work leverages the high precision of CARE and uses these results to train a model, which is not constrained by the lexicon size in its predictions. Our method also employs bootstrapping to expand the set of patterns and lexicon, similar to Agichtein and Gravano (2000) and Jones et al. (1999) but focuses on extracting affect rather than relation tuples.

3 The CARE Method

In this section, we provide a formal description of CARE for annotating the affective response of posts. Before we proceed, we note two aspects of affective responses. First, there is no formal definition for what qualifies as an affective response. In practice, we use affective responses to understand the experience that the user has when seeing a piece of content, and these responses may be both emotional and cognitive. Second, the response a user may have to a particular piece of content is clearly a very personal one. Our goal here is to predict whether a piece of content is generally likely to elicit a particular affective response. In practice, if the recommendation system has models of user interests and behavior, these would need to be combined with the affect predictions for personalized predictions.

3.1 CARE patterns and the CARE lexicon

CARE is composed of two major components: CARE patterns, regular expressions used to extract information from the comments of a post, and the CARE lexicon, a keyword-affect dictionary used to map the comment to an affect.

CARE patterns are not class or affect-specific and leverage common structure present in comments for affective response extraction. There is an unlimited number of possible CARE patterns, but we seeded the system with six CARE patterns and an additional 17 more were automatically discovered using the expansion method. In the same spirit as Hearst Patterns (Hearst, 1992), CARE patterns are tailored to extract specific relationships and rely on two sets of sub-patterns:

- Exaggerators $\{E\}$: words that intensify or exaggerate a statement, e.g., *so*, *very*, or *really*.

- Indicators $\{I\}$: words (up to 3) that exist in the CARE lexicon, which maps the indicator to a particular class. For example, 'funny' in "This is so funny" would map to *amused*.

We present the six CARE patterns below that were used to seed the system: (The symbol * (resp. $^+$) indicates that zero (resp. one) or more matches are required.) Example: *This is so amazing!*

- Demonstrative Pronouns:
  {this|that|those|these}{is|are}$^*\{E\}^*\{I\}^+$
  Example: *This is so amazing!*

- Subjective Self Pronouns:
  {i|we}{am|is|are|have|has}$^*\{E\}^*\{I\}^+$
  Example: *I am really inspired by this recipe.*

- Subjective Non-self Pronouns:
  {he|she|they}{is|are|have|has}$^*\{E\}^*\{I\}^+$
  Example: *They really make me mad.*

- Collective Nouns:
  {some people|humans|society}$\{E\}^+\{I\}^+$
  Example: *Some people are so dumb.*

- Leading Exaggerators: $\{E\}^+\{I\}^+$
  Example: *So sad to see this still happens.*

- Exclamatory Interrogatives:
  {what a|how}$\{E\}^+\{I\}^+$
  Example: *What a beautiful baby!*

Given the indicators extracted by the CARE patterns, the CARE lexicon is responsible for mapping the comment to particular affective responses. The lexicon contains 163 indicators for the 7 classes we consider (123 of which were automatically identified in the expansion process described in the next section). We also considered using other lexicons (Strapparava and Valitutti, 2004; Poria et al., 2014; Staiano and Guerini, 2014; Esuli and Sebastiani, 2006; Mohammad et al., 2013), but we found that they were lacking enough application context to be useful in our setting. Table 1 shows the affects in the CARE lexicon and corresponding definitions and example comments that would fall under each affect (or class). The classes *excited*, *angered*, *saddened*, and *scared* were chosen since they are often proposed as the four basic emotions (Wang et al., 2011; Jack et al., 2014; Gu et al., 2016; Zheng et al., 2016). The classes *adoring*, *amused*, and *approving* were established because they are particularly important in the context of social media for identifying positive content that users enjoy. Overall, a qualitative inspection indicated that these seven have minimal conceptual overlap and sufficiently broad coverage. We note,

however, that one of the benefits of the method we describe is that it is relatively easy to build a model for a new class of interest compared to the process of human annotation.

3.2 Labeling posts

Here we describe how to combine and use the two major components (CARE patterns and lexicon) at the comment-level in order to annotate the post-level affective response. The pipeline for labeling posts is shown in steps 1–3 of Figure 1 and described in detail in Appendix, Algorithm 1. We begin with reg-ex matching of CARE patterns and individual sentences of the comments. We truncate the front half of a sentence if it contains words like 'but' or 'however' because the latter half usually indicates their predominant sentiment. We also reject indicators that contain negation words such as 'never', 'not', or 'cannot' (although one could theoretically map this to the opposite affective response using Plutchik's Wheel of Emotions (Plutchik, 1980)). Note that contrary to traditional rule-based or machine-learning methods, we do not strip stop words (e.g., 'this' and 'very') because it is often crucial to the regular expression matching, and this specificity has a direct impact on the precision of the pipeline.

Given the reg-ex matches, we use the lexicon to map the indicators to the publisher affect of the comment (e.g., *excited*). It is important to note that the expressed affects of the comments should intuitively equate to the affective responses of a post. Consequently, we obtain post-level affective response labels by aggregating the comment-level labels and filtering out labels that have a support smaller than t. Specifically, a post would be labeled with the affective response a if at least t of the comments were labeled with a. In our experiments, we used a value of $t = 5$, after qualitative inspection of CARE$_{db}$, discussed in Section 4. We note, however, that it is possible for a comment to be labeled according to multiple classes if it has multiple indicators. In reality, the program should be permissive of multiple labels for a single comment, because emotions are in many cases not mutually exclusive—an individual, for example, could be experiencing both sadness and anger simultaneously.

3.3 Expanding CARE patterns/lexicon

We seeded our patterns and lexicon with a small intuitive set. We then expanded these by looking at common n-grams that appear across posts with the same label (steps A–C of Figure 1). At a high level, for a given affect a, consider the set, $comm(a)$, of all the comments on posts that were labeled a, but did not match any CARE pattern. From these comments, we extract new keywords (e.g. 'dope' for *approving* as in 'This is so dope.') for the CARE lexicon by taking the most frequent n-grams in $comm(a)$ but infrequent in $comm(b)$, where b includes all classes except a. On the other hand, the most common n-grams co-occuring with multiple classes were converted to regular expressions and then added as new CARE patterns (see Table B1 for a few examples). We added CARE patterns according to their frequency and stopped when we had sufficient data to train our models. After two expansion rounds, the set of patterns and indicators increased from 6 to 23 and 40 to 163, respectively. Counting the possible combinations of patterns and indicators, there are roughly 3500 distinct expressions. *When considering the possible 23 CARE patterns, 163 CARE lexicon indicators, and 37 exaggerators, there are a total of 130k possible instantiations of a matching comment.*

4 Evaluating CARE$_{db}$

In this section we apply our method to social media posts and validate these annotations using human evaluation (Section 4.1). Section 4.2 discusses class-wise error analysis, and in Section 4.3, we explore the alternative possibility of creating CARE$_{db}$ using a SOTA publisher-affect classifier (Demszky et al., 2020) to label the comments instead of using the CARE patterns/lexicon.

CARE$_{db}$: Our experiments use a dataset that is created from Reddit posts and comments in the pushshift.io database that were created between 2011 and 2019. We create our dataset, CARE$_{db}$, as follows. We used CARE patterns and the CARE lexicon to annotate 34 million comments from 24 million distinct posts. After filtering with a threshold of $t = 5$, we obtained annotations for 400k posts (the total number of posts that have at least 5 comments was 150 million). The low recall is expected given the specificity of CARE patterns/lexicon. We also filtered out posts that have less than 10 characters, resulting in a total of 230k posts in CARE$_{db}$. Table 1 shows the breakdown of cardinality per affective response. 195k of the posts were assigned a single label, whereas 26k (resp. 8k) were assigned two (resp. three) labels. Note that the distribution of examples per class in

AR	Definition	Example comment	Size
Adoring	Finding someone or something cute, adorable, or attractive.	*He is the cutest* thing ever.	36
Amused	Finding something funny, entertaining, or interesting.	*That was soooo funny.*	30
Approving	Expressing support, praise, admiration, or pride.	*This is really fantastic!*	102
Excited	Expressing joy, zeal, eagerness, or looking forward to something.	*Really looking forward to* this!	41
Angered	Expressing anger, revulsion, or annoyance.	*I'm so frustrated* to see this.	26
Saddened	Expressing sadness, sympathy, or disappointment.	*So sad from reading this.*	34
Scared	Expressing worry, concern, stress, anxiety, or fear.	*Extremely worried* about finals.	2

Table 1: Definition of affective responses (AR), examples of comments which would map to each affective response, and the number of posts (in thousands) per class in $CARE_{db}$. The portion of each example which would match a CARE pattern in a reg-ex search is italicized.

$CARE_{db}$ is not reflective of the distribution in the original data, because different classes have different recall rates. The $CARE_{db}$ dataset features the pushshift.io id and text of the post as well as the annotations using CARE.

4.1 Human evaluation

In our next experiment, we evaluate the labels predicted by CARE with the help of human annotators using Amazon Mechanical Turk (AMT), restricting to those who qualify as AMT Masters and have a lifetime approval rating greater than 80%. The dataset for annotation was created as follows. We sub-sampled a set of 6000 posts from $CARE_{db}$, ensuring that we have at least 500 samples from each class and asked annotators to label the affective response of each post. Annotators were encouraged to select as many as appropriate and also permitted to choose 'None of the above' as shown in Figure C1. In addition to the post, we also showed annotators up to 10 sampled comments from the post in order to provide more context. This was also done in an effort to make the comparison to CARE more fair, since CARE relies upon having access to the comments of the post. Every post was shown to three of the 91 distinct annotators. For quality control, we also verified that there no individual annotator provided answers that disagreed with the CARE labels more than 50% of the time on more than 100 posts.

We observed an average Fleiss' kappa score of 0.59, which is considered moderate agreement, the breakdown of which is shown in Table C1. Table 2 shows that the rate of agreement between the annotators and the labels proposed by the CARE method is high. For example, 94% of posts had at least one label proposed by CARE that was confirmed by 2 or more annotators, and 90% had *all* the labels confirmed. The last column measures the agreement among annotators on labels that were not suggested

by CARE, which was 53% when confirmed by 2 or more annotators. We expected this value for 'other' to be reasonably large because the CARE patterns/lexicon were designed to generate a highly precise set of labels, rather than highly comprehensive ones. However, the value is still much smaller relative to the agreement rate for the CARE labels (53% versus 94%). On average, each annotation answer contained around 1.8 labels per post (with a standard deviation of 0.9). We note that 'None of the above' was chosen less than 0.2% of the time. Table C2 and Figure C2 present annotator agreement statistics and label prevalence, respectively, broken down by class.

# Agree	Any CARE	All CARE	Other
≥ 1	98	96	82
≥ 2	94	90	53
$= 3$	80	76	24

Table 2: The rate of agreement between the annotators and the labels proposed by CARE. The first column specifies the number of annotators to be used for consensus. The rest of the columns shows for all posts, the average rate of intersection of the human labels with at least one CARE label, all CARE labels, and any label that is not a CARE label.

4.2 Error Analysis

Evaluating CARE reveals that the accuracy of CARE varies by class (Figure C2), and in particular, is lower for *amused* and *excited*. As can be seen from the interclass Spearman correlations (Figure G4) and a two-dimensional projection of the embeddings of the labeled comments (Figure G3), there appears to be non-trivial overlap amongst the classes *amused*, *excited*, and *approving*. To better understand if certain pattern or indicator matches are at fault here, we investigate the precision and recall at the pattern and lexicon level.
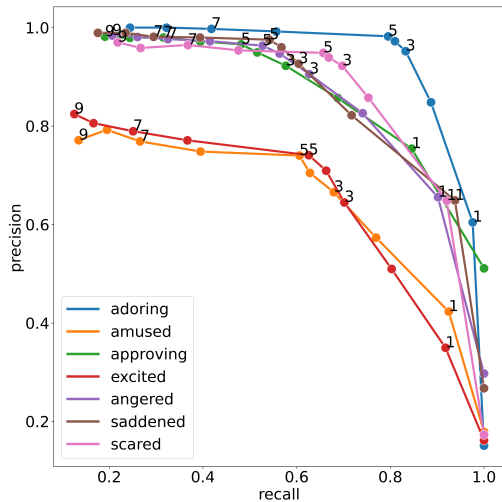
Figure 3: Precision versus recall of each class using varying thresholds (t = 0 to 9). Ground truth labels utilized are those which have at least 2 out of 3 annotator agreement. For clarity, only odd values of t are labeled.

Recall that instantiating a match for a comment involves choosing a (pattern, keyword) combination. Separating the lexicon from the patterns enables us to encode a large number of instantiated patterns parsimoniously, but some pair combinations provide a much weaker signal than others, particularly for the class *amused* (see Figure H6 for examples). Hence, for future iterations of CARE, we intend to implement a mechanism to exclude certain pattern and keyword combinations and a means for using different thresholds for each class.

Alternatively, another mechanism for accommodating these class-wise discrepancies in performance is by tuning for each class an optimal threshold t (i.e., the number of matched comments we need to see in order to reliably predict a label). Figure 3 shows how the precision and recall of each class varies according to different threshold values. To achieve precision and recall greater than 0.7, a threshold of 1 actually seems viable for most classes, while for *amused* and *excited* a threshold of at least 3 is needed. In fact, for most of the classes, using thresholds larger than 3 has negligible impact on the precision score, but does reduce the recall.

4.3 Can we leverage emotion classification?

Recall, steps 1 and 2 of Figure 1 uses the CARE patterns and lexicon to label the publisher affect of the comments. Conceivably, this could have been done instead by using a SOTA emotion classifier such as the GoEmotions classifier (Demszky et al., 2020), which is trained specifically to predict the publisher affect of Reddit comments. Here we show that our method for labeling the publisher affect of comments performs comparatively better. Let us define the method $CARE^G$, a modified version of the CARE method where steps 1 and 2 are replaced with labels using the GoEmotions classifer. We apply $CARE^G$ to our human annotated dataset (Section 4.1) by first applying the GoEmotions classifier to all comments of the posts. These GoEmotion labels are then mapped to our taxonomy in Table 1 using the mapping defined in Table 3, which is based on the grouping of emotions at the Ekman level used in Demszky et al. (2020). We then, as usual, aggregate and filter post labels according to a threshold t.

$CARE^G$ (Table F4) shows a relative decrease of 12.9% and 18.0% in the rate of annotator agreement with any and all labels, respectively, compared to that of CARE. These decreases hold even when partitioning on each individual class. The comparatively lower performance of $CARE^G$ is most likely due to the low F1-scores (<0.4) of the GoEmotions classifer for nearly half of the 28 classes, as reported in the original work Demszky et al. (2020, Table 4). It is also important to note that in addition to demonstrating higher precision, CARE patterns and lexicon are valuable because they do not require human annotated data, unlike GoEmotions. It may, however, be useful to leverage multiple emotion detection approaches. Section F discusses a potential ensembling strategy for this.

To validate the mapping in Table 3, we applied steps 1 and 2 of CARE to the GoEmotions dataset (see Section E), and computed the rate of agreement among the labels in our defined mapping. We find this rate of agreement to be high (87.3% overall). Note, we perform this equivalence at the publisher affect level, because as discussed before, the affective response and publisher affect are not always equivalent. In addition to prior work (Dwivedi-Yu et al., 2022), Section D presents experiments that indicate that affective response and publisher affect labels intersect only 44% of the time.

5 Predicting affective response for posts without comments

In this section we describe CARE-BERT, a multi-label affective response classifier that is trained only on the post-level text and annotations in $CARE_{db}$. Such a model is important in order to

AR	GoEmotion label	% agree
Amused	Amusement	79.8
Approving	Admiration, approval	89.3
Excited	Joy	81.3
Angered	Anger, Annoyance, Disgust	93.3
Saddened	Disappointment, sadness	90.9
Scared	Fear, nervousness	84.9

Table 3: CARE to GoEmotions mapping. The last column summarizes the rate at which the mapping holds. The average across all datapoints was 87.3%.

make predictions early in the life of the post and in cases where the comments may not exist or may not match any CARE patterns or keywords. Note that *the model is not given the comments* text and is therefore not restricted to the CARE pattern/lexicon semantics. In section 5.2, we describe how CARE-BERT can be further fine-tuned for related tasks like emotion detection.

5.1 Creating and evaluating CARE-BERT

We train CARE-BERT with the CARE labels in CARE$_{db}$, using the pre-trained model bert-base-uncased from the Huggingface library (Wolf et al., 2020). We use a max length of 512 and we add a dropout layer with a rate of 0.3 and a dense layer to allow for multi-label classification. We used an Adam optimizer with a learning rate of 5e-5, a batch size of 16, and 5 epochs. We used a train/validation/test split of 80/10/10%. See Section I for other settings we explored.

The evaluation on the human-annotated set (held out from training) is shown in Table 4. We use labels with support from all annotators as ground truth. The classes of lowest prevalence, such as *scared*, had the poorest results, while the more frequent classes (*adoring*, *approving*, *saddened*) had the highest results. To put these results in perspective, we use the mapping in Table 3 and compare with the numbers from Demszky et al. (2020). Note, the comparison is *not* for the same dataset—our results pertain to predicting on the post whereas GoEmotions predicts the comments. Still, CARE-BERT demonstrates a 35% improvement in the overall micro-averaged F1-score.

CARE vs. CARE-BERT: Compared to the human annotators and CARE, CARE-BERT is disadvantaged by not having access to the comments. We use human annotated set of CARE$_{db}$ and find that 0.89 of the CARE labels are also proposed by human annotators, while this value is 0.72 for

CARE-BERT (Table J6). In Table J7 we display select examples that may illustrate reasons for this discrepancy. Firstly, one of the challenges that CARE-BERT faces is that there may not be sufficient context in the post alone. In the example "Who is this LIRIK guy, and why does he have 50K subscribers" it is challenging to predict that some people find the subject adorable without additional context. Relatedly, the conversation that the post initiates can be challenging to foresee. The last example reads "AskReddit: Imagine the last thing you ate has been made illegal. What would that be?" In some cases, commenters just ate something they didn't like and are therefore content with the premise. In other cases, commenters just ate something they very much enjoy and are saddened by the hypothetical. Our results show that this is not particular to 'AskReddit' posts, and given these challenges, it is reasonable that the CARE method provides more reliable labels.

Affect	P	R	F1	GoEmotions F1
Adoring	0.73	0.66	0.70	-
Amused	0.63	0.54	0.60	**0.80**
Approving	0.73	0.72	**0.75**	0.53
Excited	0.58	0.52	**0.58**	0.51
Angered	0.70	0.61	**0.69**	0.40
Saddened	0.78	0.62	**0.73**	0.39
Scared	0.68	0.3	0.47	**0.54**
micro-avg	0.70	0.68	**0.69**	0.51
macro-avg	0.69	0.62	**0.65**	0.53
stdev	0.06	0.12	0.09	0.14

Table 4: Precision (P), recall (R), and F1 of CARE-BERT using CARE$_{db}$ on the post text of the human-annotated set and F1-scores of the GoEmotions classifier from Demszky et al. (2020) on comments.

5.2 Transfer learning to emotion detection

We now demonstrate that CARE-BERT is also useful for pre-training of another related task in a setting with limited annotated data. We consider transfer learning to the ISEAR Dataset (Scherer and Wallbott, 1994), which is a collection of 7666 statements from a diverse set of 3000 individuals labeled according to six categories (anger, disgust, fear, guilt, joy, sadness, and shame). The labels pertain to the *publisher affect* and not affective response, as considered in this work. Our experiment explores transfer learning to predict the labels in the ISEAR dataset using an additional drop-out layer of 0.3 and a dense layer.

Our experiments follow closely to that of Dem-

szky et al. (2020) and uses different training set sizes (500, 1000, 2000, 4000, and 6000) for 10 different train-test splits. We plot the average and standard deviation in the F1-scores across these 10 splits in Figure 4. We compare four different fine-tuning setups: the first two are trained using CARE-BERT and then fine-tuned on the benchmark dataset, one with no parameter freezing (no_freeze), and one with all layers but the last two frozen (freeze). The third setup is similar to CARE-BERT (no_freeze) but is trained on GoEmotions rather than CARE$_{db}$. The last setup is the bert-base-uncased model trained only on ISEAR, where all setups use the same architecture and hyperparameters as discussed in Section 5.

Our values differ slightly from that cited in Demszky et al. (2020) due to the small differences in architecture and hyperparameters. However, the overall results corroborate that of Demszky et al. (2020) in that models with additional pre-training perform better than the baseline (no additional pre-training) for limited sample sizes. From Figure 4, it is apparent that CARE-BERT and the model built from GoEmotions perform essentially on par in these transfer learning experiments in spite of the fact that CARE-BERT does not utilize human annotations. It is also worth noting that GoEmotions and the ISEAR dataset address the same task (emotion detection) while CARE-BERT predicts affective response. The comparable performance of CARE-BERT with the GoEmotions models demonstrates the utility of CARE-BERT for other tasks with limited data and the promise of CARE as a means of reliable unsupervised labeling.

6 Conclusion

We described a method for extracting training data for models that predict the affective responses of a post on social media. CARE is an efficient, accurate, and scalable way of collecting unsupervised labels and can be extended to new classes. Using CARE, we created CARE$_{db}$, a large dataset which can be used for affective response detection and other related tasks, as demonstrated by the competitive performance of CARE-BERT to similar BERT-based models in emotion detection. We release the annotations and models in the hopes that this will unlock future research.

In particular, there are two main cases in which CARE can be improved upon: (1) when there does not exist a set of common phrases that are indica-
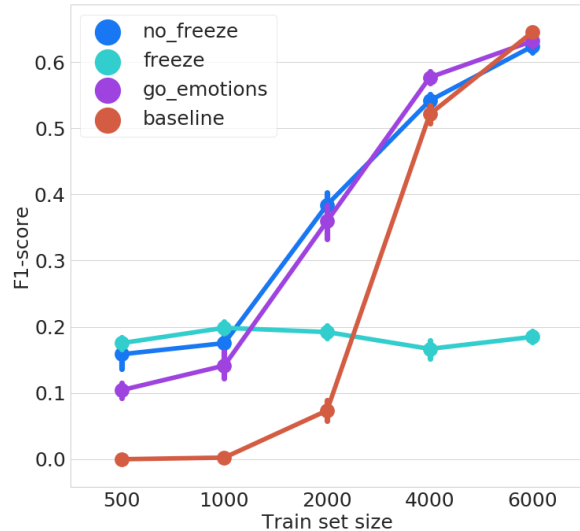

Figure 4: The F1-score of each model using varying training set sizes of the ISEAR dataset. The light blue line refers to using CARE-BERT, but with freezing all parameters except in the last layer. The dark blue is the same but without freezing. Lastly, the purple line refers to the same architecture as CARE-BERT (no freezing) but trained on GoEmotions instead of CARE$_{db}$, and the red line is trained only on the ISEAR dataset using a bert-base-uncased model with the same hyperparameters.

tive of an affect, and (2) when an indicator maps to multiple affects. In the latter case, there is still partial information that can be gleaned from the labels. In addition to developing methods for the above cases, future work also includes incorporating emojis, negations, and punctuation, extending to new classes, or possibly using embedding similarity rather than exact match for the CARE patterns. Finally, we also plan to investigate the use of CARE for predicting the affective response to images as well as multi-modal content such as memes.

References

Francisca Adoma Acheampong, Chen Wenyu, and Henry Nunoo-Mensah. 2020. Text-based emotion detection: Advances, challenges, and opportunities. *Engineering Reports*, page e12189.

Eugene Agichtein and Luis Gravano. 2000. Snowball: Extracting relations from large plain-text collections. In *Proceedings of the fifth ACM conference on Digital libraries*, pages 85–94.

Sheeraz Akram, Shariq Hussain, Ibrahima Kalil Toure, Shunkun Yang, and Humza Jalal. 2020. Choseamobile: A web-based recommendation system for mobile phone products. *Journal of Internet Technology*, 21(4):1003–1011.

Nourah Alswaidan and Mohamed Menai. 2020. A survey of state-of-the-art approaches for emotion recognition in text. *Knowledge and Information Systems*, 62.

Sergio Angelastro, B Carolis, and Stefano Ferilli. 2019. Predicting user preference in pairwise comparisons based on emotions and gaze. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pages 253–261. Springer.

Olga Artemenko, Volodymyr Pasichnyk, Nataliia Kunanets, and Khrystyna Shunevych. 2020. Using sentiment text analysis of user reviews in social media for e-tourism mobile recommender systems. In *COLINS*, pages 259–271.

Dr. Muhammad Asghar, Aurangzeb Khan, Afsana Bibi, Fazal Kundi, and Hussain Ahmad. 2017. Sentence-level emotion detection framework using rule-based classification. *Cognitive Computation*, 9:1–27.

Shenghua Bao, Shengliang Xu, Li Zhang, Rong Yan, Zhong Su, Dingyi Han, and Yong Yu. 2012. Mining social emotions from affective text. *IEEE Transactions on Knowledge and Data Engineering*, 24(9):1658–1670.

Francesco Barbieri, Jose Camacho-Collados, Francesco Ronzano, Luis Espinosa-Anke, Miguel Ballesteros, Valerio Basile, Viviana Patti, and Horacio Saggion. 2018. SemEval 2018 task 2: Multilingual emoji prediction. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 24–33, New Orleans, Louisiana. Association for Computational Linguistics.

Christos Baziotis, Nikos Pelekis, and Christos Doulkeridis. 2017. DataStories at SemEval-2017 task 4: Deep LSTM with attention for message-level and topic-based sentiment analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 747–754, Vancouver, Canada. Association for Computational Linguistics.

Yan-Ying Chen, Tao Chen, Winston H Hsu, Hong-Yuan Mark Liao, and Shih-Fu Chang. 2014. Predicting viewer affective comments based on image content in social media. In *proceedings of international conference on multimedia retrieval*, pages 233–240.

Jérémie Clos, Anil Bandhakavi, Nirmalie Wiratunga, and Guillaume Cabanac. 2017. Predicting emotional reaction in social networks. In *European Conference on Information Retrieval*, pages 527–533. Springer.

Michael AA Cox and Trevor F Cox. 2008. Multidimensional scaling. In *Handbook of data visualization*, pages 315–347. Springer.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

Jane Dwivedi-Yu, Yi-Chia Wang, Lijing Qin, Cristian Canton-Ferrer, and Alon Y Halevy. 2022. Affective signals in a social media recommender system. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2831–2841.

Andrea Esuli and Fabrizio Sebastiani. 2006. SENTI-WORDNET: A publicly available lexical resource for opinion mining. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1615–1625, Copenhagen, Denmark. Association for Computational Linguistics.

Karën Fort, Gilles Adda, and Kevin Bretonnel Cohen. 2011. Amazon mechanical turk: Gold mine or coal mine? *Computational Linguistics*, pages 413–420.

Lisa Graziani, Stefano Melacci, and Marco Gori. 2019. Jointly learning to detect emotions and predict facebook reactions. In *International Conference on Artificial Neural Networks*, pages 185–197. Springer.

Simeng Gu, Wei Wang, Fushun Wang, and Jason H Huang. 2016. Neuromodulator and emotion biomarker for stress induced mental disorders. *Neural plasticity*, 2016.

Maryam Hasan, Emmanuel Agu, and Elke Rundensteiner. 2014. Using hashtags as labels for supervised learning of emotions in twitter messages. In *ACM SIGKDD workshop on health informatics, New York, USA*.

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *COLING 1992 Volume 2: The 15th International Conference on Computational Linguistics*.

Chenyang Huang, Amine Trabelsi, and Osmar Zaïane. 2019a. ANA at SemEval-2019 task 3: Contextual emotion detection in conversations through hierarchical LSTMs and BERT. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 49–53, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Yen-Hao Huang, Ssu-Rui Lee, Mau-Yun Ma, Yi-Hsin Chen, Ya-Wen Yu, and Yi-Shin Chen. 2019b. Emotionx-idea: Emotion bert–an affectional model for conversation. *arXiv preprint arXiv:1908.06264*.

Rachael E Jack, Oliver GB Garrod, and Philippe G Schyns. 2014. Dynamic facial expressions of emotion transmit an evolving hierarchy of signals over time. *Current biology*, 24(2):187–192.

Rosie Jones, Andrew McCallum, Kamal Nigam, and Ellen Riloff. 1999. Bootstrapping for text learning tasks. In *IJCAI-99 Workshop on Text Mining: Foundations, Techniques and Applications*, volume 1. Citeseer.

Brendan Jou, Subhabrata Bhattacharya, and Shih-Fu Chang. 2014. Predicting viewer perceived emotions in animated gifs. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 213–216.

Phil Katz, Matt Singleton, and Richard Wicentowski. 2007. SWAT-MP:the SemEval-2007 systems for task 5 and task 14. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 308–313, Prague, Czech Republic. Association for Computational Linguistics.

Florian Krebs, Bruno Lubascher, Tobias Moers, Pieter Schaap, and Gerasimos Spanakis. 2017. Social emotion mining techniques for facebook posts reaction prediction. *arXiv preprint arXiv:1712.03249*.

Jingsheng Lei, Yanghui Rao, Qing Li, Xiaojun Quan, and Liu Wenyin. 2014. Towards building a social emotion detection system for online news. *Future Generation Computer Systems*, 37:438 – 448. Special Section: Innovative Methods and Algorithms for Advanced Data-Intensive Computing Special Section: Semantics, Intelligent processing and services for big data Special Section: Advances in Data-Intensive Modelling and Simulation Special Section: Hybrid Intelligence for Growing Internet and its Applications.

Kevin Hsin-Yih Lin, Changhua Yang, and Hsin-Hsi Chen. 2008. Emotion classification of online news articles from the reader's perspective. In *2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, volume 1, pages 220–226. IEEE.

Chunling Ma, Helmut Prendinger, and Mitsuru Ishizuka. 2005. Emotion estimation and reasoning based on affective textual interaction. In *Proceedings of the First International Conference on Affective Computing and Intelligent Interaction*, ACII'05, page 622–628, Berlin, Heidelberg. Springer-Verlag.

Saif M Mohammad and Svetlana Kiritchenko. 2015. Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence*, 31(2):301–326.

Saif M Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. *arXiv preprint arXiv:1308.6242*.

Gabriele Paolacci and Jesse Chandler. 2014. Inside the turk: Understanding mechanical turk as a participant pool. *Current directions in psychological science*, 23(3):184–188.

Robert Plutchik. 1980. Chapter 1 - a general psycho-evolutionary theory of emotion. In Robert Plutchik and Henry Kellerman, editors, *Theories of Emotion*, pages 3 – 33. Academic Press.

Chris Pool and Malvina Nissim. 2016. Distant supervision for emotion detection using Facebook reactions. In *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES)*, pages 30–39, Osaka, Japan. The COLING 2016 Organizing Committee.

Soujanya Poria, Alexander Gelbukh, Erik Cambria, Amir Hussain, and Guang-Bin Huang. 2014. Emosenticspace: A novel framework for affective common-sense reasoning. *Knowledge-Based Systems*, 69:108 – 123.

Bin Tareaf Raad, Berger Philipp, Hennig Patrick, and Meinel Christoph. 2018. Aseds: Towards automatic social emotion detection system using facebook reactions. In *2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, pages 860–866. IEEE.

Yanghui Rao, Qing Li, Liu Wenyin, Qingyuan Wu, and Xiaojun Quan. 2014. Affective topic model for social emotion detection. *Neural Networks*, 58:29 – 37. Special Issue on "Affective Neural Networks and Cognitive Learning Systems for Big Data Analysis".

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Renata L Rosa, Demsteneso Z Rodriguez, and Graça Bressan. 2015. Music recommendation system based on user's sentiments extracted from social networks. *IEEE Transactions on Consumer Electronics*, 61(3):359–367.

Renata Lopes Rosa, Gisele Maria Schwartz, Wilson Vicente Ruggiero, and Demóstenes Zegarra Rodríguez. 2018. A knowledge-based recommendation system that includes sentiment analysis and deep learning. *IEEE Transactions on Industrial Informatics*, 15(4):2124–2135.

Klaus R. Scherer and Harald Wallbott. 1994. Evidence for universality and cultural variation of differential emotion response patterning. *Journal of personality and social psychology*, 66(2):310.

Valentina Sintsova and Pearl Pu. 2016. Dystemo: Distant supervision method for multi-category emotion recognition in tweets. *ACM Trans. Intell. Syst. Technol.*, 8(1).

Jacopo Staiano and Marco Guerini. 2014. Depeche mood: a lexicon for emotion analysis from crowd annotated news. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 427–433, Baltimore, Maryland. Association for Computational Linguistics.

Luke Stark and Jesse Hoey. 2021. The ethics of emotion in artificial intelligence systems. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 782–793.

Carlo Strapparava and Rada Mihalcea. 2007. Semeval-2007 task 14: Affective text. *Association for Computational Linguistics*, pages 308–313.

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).

Jared Suttles and Nancy Ide. 2013. Distant supervision for emotion classification with discrete binary values. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 121–136. Springer.

Jianhua Tao. 2004. Context based emotion detection from text input. In *Eighth International Conference on Spoken Language Processing*.

Luis Tapia. *Clown on black background*. Shutterstock.

Marko Tkalčič, Nima Maleki, Matevž Pesek, Mehdi Elahi, Francesco Ricci, and Matija Marolt. 2017. A research tool for user preferences elicitation with facial expressions. In *Proceedings of the eleventh acm conference on recommender systems*, pages 353–354.

Marko Tkalčič, Nima Maleki, Matevž Pesek, Mehdi Elahi, Francesco Ricci, and Matija Marolt. 2019. Prediction of music pairwise preferences from facial expressions. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 150–159.

Kaiyu Wang, Yanmeng Guo, Fei Wang, and Zuoren Wang. 2011. Drosophila trpa channel painless inhibits male–male courtship behavior through modulating olfactory sensation. *PLoS One*, 6(11):e25890.

Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P Sheth. 2012. Harnessing twitter "big data" for automatic emotion identification. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, pages 587–592. IEEE.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zheng Zheng, Simeng Gu, Yu Lei, Shanshan Lu, Wei Wang, Yang Li, and Fushun Wang. 2016. Safety needs mediate stressful events induced mental disorders. *Neural plasticity*, 2016.

A Broader Impact

Any work that touches upon emotion recognition or recognizing affective response needs to ensure that it is sensitive to the various ways of expressing affect in different cultures and individuals. Clearly, applying the ideas described in this paper in a production setting would have to first test for cultural biases. To make "broad assumptions about emotional universalism [would be] not just unwise, but actively deleterious" to the general community (Stark and Hoey, 2021). We also note that emotion recognition methods belong to a taxonomy of conceptual models for emotion (such as that of Stark and Hoey (2021) and these "paradigms for human emotions [...] should [not] be taken naively ground truth."

Before being put in production, the method would also need to be re-evaluated when applied to a new domain to ensure reliable performance in order to prevent unintended consequences. Additionally, our work in detecting affective response is intended for understanding content, not the emotional state of individuals. This work is intended to identify or recommend content, which aligns with the user's preferences. This work should not be used for ill-intended purposes such as purposefully recommending particular content to manipulate a user's perception or preferences.

B Details on expanding CARE

n-gram	frequency	class
adorable	9000	Adoring
gorgeous	8422	Adoring
fantastic	7796	Approving
interesting	5742	Amused
sorry for your	5202	Saddened
brilliant	4205	Approving
fake	2568	Angered
sorry to hear	2323	Saddened
why i hate	1125	Angered
i feel like	293	pattern
you are a	207	pattern
this is the	173	pattern
this made me	110	pattern
he is so	102	pattern

Table B1: Examples of n-grams resulting from `GetNgrams` in Algorithm 1 and steps B1 and B2 of Figure 1. The n-grams above the middle line are added to the lexicon under the specific class listed while the n-grams below are used for further expansion of CARE patterns after translating to reg-ex format manually.

Algorithm 1 on page 18 presents pseudo-code for the process of labeling posts and expanding CARE patterns and the CARE lexicon. Table B1 presents example results from the expansion process.

C Annotation details

Figure C1: Interface for crowdsourcing process using Amazon Mechanical Turk. Three distinct annotators were used to annotate each post. Annotators were told an affective response is an emotion or cognitive response to the post and the definitions and examples in Table 1 were shown to them.

AR	% w/ support	Avg support	Fleiss' kappa
Adoring	99.2	2.8	0.78
Amused	93.2	2.1	0.43
Approving	98.8	2.8	0.51
Excited	83.6	2.1	0.58
Angered	99.4	2.8	0.59
Saddened	99.6	2.9	0.61
Scared	98.8	2.6	0.64
Average	**96.1**	**2.6**	**0.59**

Table C2: The percent of CARE-labeled examples (maximum of 100) with agreement from at least one labeler by class and of those examples, the average number of annotator agreement (maximum of 3). The third column shows the Fleiss' kappa, which was computed for class a based on the presence and absence of label a by each annotator for a given post. The bottom row is the average over all classes.

The annotators were paid a competitive wage in order to temper the effects of the ethical and sampling limitations and concerns as described in Fort et al. (2011) and Paolacci and Chandler (2014). Figure C1 shows the interface used for crowdsourcing human annotations for evaluating CARE patterns. To better understand annotation results for each class, we present Table C2, which shows annotator agreement statistics broken down by class.

We also computed Fleiss' kappa for each class, where a value between 0.41-0.60 is generally considered moderate agreement and a value between 0.61-0.80 is substantial agreement. As can be seen, classes such as *adoring* have high average annotator support and Fleiss' kappa while others like *amused* have low average annotator support and Fleiss' kappa, an observation that aligns with the findings in Section 4.2.
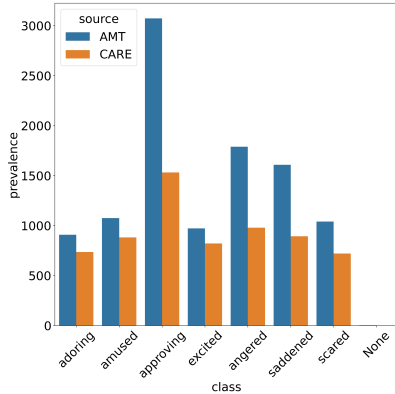

Figure C2: Prevalence of class labels according to annotations from AMT on which at least two annotators agree upon (blue) and according to CARE (orange). The prevalence of *approving* was much higher from AMT, likely due to a large perceived overlap in the definitions of *approving* and other classes such *excited*.

D Are affective response and publisher affect the same?

The GoEmotions dataset and classifier target the publisher affect (of comments), whereas CARE-BERT and CARE target the affective response (of posts). In an effort to study the correlation between affective response and publisher affect, we compare the following sets of labels: 1) human annotations of GoEmotion and the predicted affective responses using CARE-BERT applied to GoEmotions and 2) CARE labels for posts in $CARE_{db}$ and the predicted publisher affects using the GoEmotions classifier applied to $CARE_{db}$. Specifically, for every annotated label (i.e., not from a classifier) we count the percentage of the time where there is intersection with the set of predicted labels (i.e., from a classifier).

The results of these experiments are shown in Table D3, broken down according to the class of the annotated label. Overall, the percentage of affective response and publisher affect label agreement (44%) is moderate but seems to indicate that the affective response detection and emotion detection

are not necessarily the same problem, in particular for *scared* and *approving*. The classes *approving*, *excited*, and *angered* have a large variance between the two datasets, where the first (Table D3, second column) uses comments and the second (Table D3, third column) uses posts. This could be due to the classification errors (either by GoEmotions or by CARE-BERT) or due to the type of the text (comment or post). More research and data collection is needed to understand the relationship between affective response and publisher affect.

AR	GoEmotions	$CARE_{db}$	Average
Amused	63	54	59
Approving	8	47	28
Excited	52	24	38
Angered	4	74	39
Saddened	60	62	61
Scared	44	34	39
Average	**39**	**49**	**44**

Table D3: Rate of intersection between affective response and publisher affect labels. The first column denotes the class. The second column denotes the percent of the time an annotated label in GoEmotions exists in the set of predicted labels by CARE-BERT when applied to the GoEmotions dataset. The third column denotes the percent of the time an annotated label in CARE-BERT exists in the set of predicted labels by the GoEmotions classifier when applied to $CARE_{db}$. The last column is the row-wise average.

E Using CARE patterns/lexicon to predict publisher affect in GoEmotions

The GoEmotions dataset (Demszky et al., 2020) is a collection of 58k Reddit comments labeled according to the publisher affect from a taxonomy of 28 emotions. There exists a natural mapping from 6 of our classes to those of GoEmotions (the exception being *adoring*) based on the definitions alone. Hence, applying CARE patterns/lexicon to the GoEmotions dataset presents another way of validating the quality of steps 1 and 2 of CARE. The number of examples in GoEmotions with labels belonging to these 6 classes was 21.0k and the number of comments that were labeled by CARE patterns/lexicon was 1259. Table 3 compares the human annotations in the GoEmotions dataset with the labels that CARE patterns/lexicon assigned to the comments and shows that they have a high degree of agreement.

While the low recall is certainly a limitation of CARE patterns and lexicon when applied to a specific small dataset, we emphasize that the pri-

mary intention of CARE patterns is to generate a labeled dataset in an unsupervised manner, so one can start training classifiers for that affective response. Given the abundance of freely available unlabeled data (e.g., on Reddit, Twitter), recall is not a problem in practice. In the next section and in Section 4.3, however, we discuss how existing emotion classifiers, such as the GoEmotions classifier (Demszky et al., 2020) can also be leveraged in the CARE method.

F CARE and CARE^G evaluation details

CARE^G refers to the CARE method, where steps 1 and 2 of Figure 1 use the GoEmotions classifier instead of CARE patterns. To evaluate how CARE and CARE^G compares, we use the same human-labeled dataset described in Section 4.1 and applied the GoEmotions classifier to all the comments belonging to these posts (72k comments). We then mapped the predicted GoEmotion labels to CARE pattern labels using the mapping in Table 3. GoEmotion and CARE labels not in the mapping are excluded from this analysis.

Threshold	Any CARE^G	All CARE^G	Other
$t = 1$	95	34	25
$t = 2$	91	61	42
$t = 3$	87	71	51
$t = 4$	81	73	57
$t = 5$	73	67	62
$t = 6$	58	56	70
$t = 7$	47	45	76
$t = 8$	38	37	81
$t = 9$	30	29	84
$t = 10$	24	23	88
max	89	89	60
CARE	93	89	54
ensemble	94	83	49

Table F4: The rate of intersection between labels agreed upon by at least two annotators and the labels proposed by CARE^G. The first column indicates the threshold t used in CARE^G. Using annotations agreed upon by at least two annotators, the rest of the columns show the rate of agreement with at least one predicted label, all predicted labels, and any human-annotated label that was not predicted. The row labeled 'max' refers to choosing the comment-level label with the highest frequency for each post. For context, the results for CARE using $t = 5$ are shown in the penultimate row. The last row presents results from combining the CARE pattern labels and the GoEmotion labels using $t = 4$.

The same metrics for ≥ 2 annotator agreement in Table 2 are shown in Table F4 for multiple thresholds and for all classes, excluding *adoring*. CARE

labels consistently demonstrate higher agreement with human annotations than those of CARE^G. The last row of Table F4 shows results for an ensembling approach where steps 1 and 2 use labels from both CARE patterns in addition to the labels from the GoEmotions classifier, where the former uses $t = 5$ and the latter uses $t = 4$ in step 3 (optimal values for each approach, respectively). This ensembling approach does reasonably well and can be used to include classes in the GoEmotions taxonomy that do not exist in the taxonomy of Table 1. Given other emotion classifiers, one could potentially include those as well.

G Multi-dimensional scaling pairwise plots

Figure G3: The two-dimensional projection (using MDS) of sentence embeddings of comments suggests that the CARE-based predictions correspond to similarity in the embedding space. Colors correspond to the labels given by CARE labeling, which were not given to the embedding model or the MDS.

We visualize the degree of overlap between the sentence embeddings (using Sentence-Bert (Reimers and Gurevych, 2019)) of 100 comments in CARE$_{db}$ for each class. We then use multi-dimensional scaling or MDS (Cox and Cox, 2008) to map the embeddings to the same two-dimensional space using euclidean distance as the similarity metric, as shown in Figure G3 and Figure G5. Note that the MDS process does not use the class labels. As can be seen, there is substantial overlap between *amused* and other classes, as well as between *excited* and *approving*. Figure G4 shows the Spearman correlation between

each class and a hierarchical clustering using the AMT-annotated dataset, and corroborates that *approving* and *excited* indeed do have the highest degree of correlation. Given that the average number of human annotations per post was 1.8 (Section 4.1), it is likely that a portion of this overlap can be attributed to the multi-label nature of the problem as well as the moderate correlation between certain classes such as *excited* and *approving* (Figure G4). See Figure G5 for plots of multi-dimensional scaling for every pair of classes, as referenced in Section 4.2.
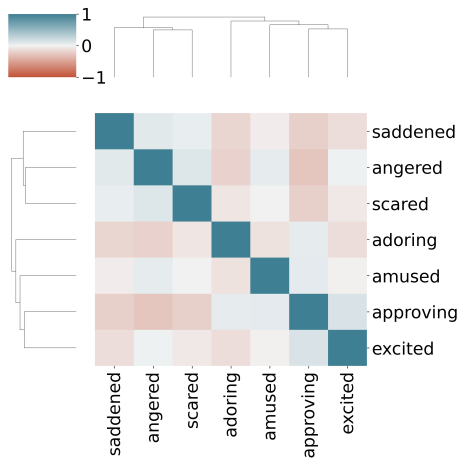

Figure G4: Pairwise Spearman correlation between each pair of classes, computed using the degree of annotator support for each class given a post. The dendrogram represents a hierarchical clustering of the data, correctly capturing the distinction between positive and negative classes.

H   Pattern match analysis

To investigate why higher thresholds would be needed for certain classes, we analyze the CARE patterns and lexicon at the class level.

Let us define a match as a tuple containing the pattern name and the word or phrase which maps the comment to an affect according to the CARE lexicon. We could also consider exaggerators in our analysis but here we assume a negligible effect on differentiating reliability. We previously assumed that each instantiated match should have the same weight of 1, but this may not be appropriate considering that some patterns or words may be more reliable.

As can be seen in Figure H6, there are some cases in which the keyword in general seems to have a high false positive rate (e.g., happy) and in

other cases it appears the erroneous combination of a particular pattern and keyword can lead to high false positive rates. For example, while the match '(so very, funny)' has a low false positive rate of 0.2, '(I, funny)' has a much higher false positive rate of 0.57, which intuitively makes since 'I'm funny' does not indicate being amused. We also investigated whether individual patterns are prone to higher false positive rates, which does not seem to be the case. For future iterations of CARE, one could also use the true positive rate as the weight of a match to obtain a weighted sum when aggregating over comments to label a post.
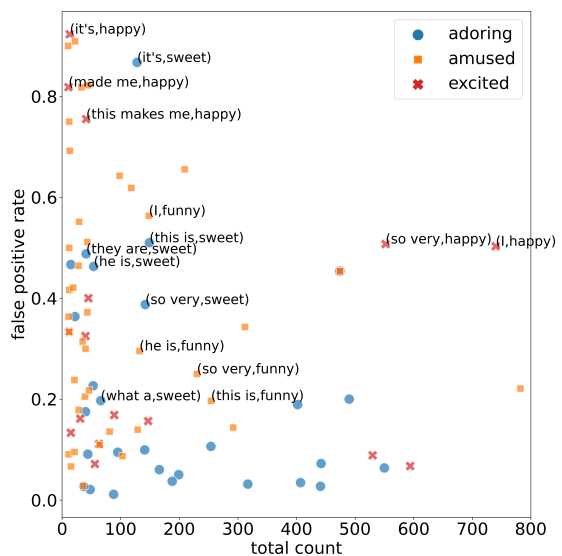

Figure H6: Scatter plot of the total frequency of a match versus its false positive rate. Ground truth labels used here are those from AMT and agreed upon by at least 2 annotators. For clarity, a match is shown only if its total count was 10 or more and if it belongs to one of the three classes (*adoring*, *amused*, and *excited*). Only those which contain the keywords 'sweet' (*adoring*), 'funny' (*amused*), and 'happy' (*excited*) are labeled.

I   Modeling details

AR	Precision	Recall	F1
Positive	0.95	0.95	0.94
Negative	0.77	0.77	0.78
micro-avg	**0.89**	**0.91**	**0.90**
macro-avg	**0.86**	**0.86**	**0.86**
stdev	**0.10**	**0.13**	**0.11**

Table I5: Accuracy of CARE-BERT for the two-class case: POSITIVE versus NEGATIVE. Note that amused, excited, adoring, and approving were mapped to positive and angered, saddened, and scared were mapped to negative.

We began with the hyper-parameter settings in Demszky et al. (2020) and explored other hyper-parameter settings (batch sizes [16, 32, 64], max length [64, 256, 512], drop out rate [0.3, 0.5, 0.7], epochs [2-10]) but found minimal improvements in the F1-score, as computed by the `scikit-learn` package in python. Running this on two Tesla P100-SXM2-16GB GPUs took roughly 19 hours. We also experimented with higher thresholds for the parameter t (see Section 3.2) but saw marginal improvements, if any.

We developed two versions of CARE-BERT: one using the classes in Table 1, and a simpler one using only the classes POSITIVE, and NEGATIVE. The first four rows in Table 1 are considered positive while the last three are negative, the results of which are featured in Table I5. Naturally, the two-class model that blurs the differences between classes with the same valence has higher results.

J   Modeling Analysis

	Human	CARE	CARE-BERT
Human	1.0	0.55	0.51
CARE	0.89	1.0	0.72
CARE-BERT	0.72	0.62	1.0

Table J6: Percentage of agreement between annotation schemes. Each entry corresponds to the percentage of all labels the annotation scheme along the row agrees with the annotation scheme along the column.

Algorithm 1: Algorithm for producing candidates for new CARE patterns and indicators in the CARE lexicon. Algorithm uses three hyperparameters t (the minimum number of comments to label a post), $f\_lexicon$ (the minimum frequency of a ngram to be added to the lexicon), and $f\_pattern$ (the minimum frequency of an ngram to be a candidate pattern) which was set to 5, 1000, and 100, respectively. The resulting list of candidate patterns needs to be manually converted into a regular expression matching the structure outlined in Section 3.1.

Data: C: set of comments, P: set of corresponding posts, L: dictionary of keywords to class (CARE lexicon), D: list of non-class-specific regular expressions (CARE patterns)

1   $lexicon\_candidates \leftarrow [], pattern\_candidates \leftarrow []$
2   $labeled\_posts \leftarrow$ LabelPosts $(C, P, L, D), ngrams \leftarrow$ GetNgrams (labeled_posts, C)
3   **for** *a in all classes* **do**
4   // Add an ngram as a lexicon candidate if it is exclusively in high frequency with class a
5   **for** *ngram in ngrams[a]* **do**
6   **if** *frequency of ngram in ngrams[a]* $\geq f\_lexicon$ **then**
7   **for** *b in all classes where* $b \neq a$ **do**
8   **if** *ngram in ngrams[b] and frequency of ngram in ngrams[b]* $\geq f\_lexicon$ **then**
9   Break and continue to new n-gram
10   Append ngram to $lexicon\_candidates$, if not added already
11   // Add an ngram as a pattern candidate if in high enough frequency and present in another class
12   **for** *ngram in ngrams[a]* **do**
13   **if** *total freq. of ngram in ngrams* $\geq f\_pattern$ *and ngram in ngrams[b] for* $b \neq a$ **then**
14   Append $ngram$ to $pattern\_candidates$, if not added already

Result: $lexicon\_candidates, pattern\_candidates$

15
16 **Function** LabelPosts(C, P, L, D)**:**
17    $labeled\_comments \leftarrow \{\}, labeled\_posts \leftarrow \{\}$
18   // For each comment, apply reg-ex and map indicator to affect using the lexicon
19   **for** *c in C* **do**
20   **if** *indicator is non-empty after reg-ex matching and in lexicon* **then**
21   Append c to $labeled\_comments[L[indicator]]$
22   // For each post, aggregate comment labels to label post
23   **for** *p in P* **do**
24   **for** *a in all classes* **do**
25   **if** *number of comments belonging to post p and labeled as class a* $\geq t$ **then**
26   Append p to $labeled\_posts[a]$
27   **return** $labeled\_posts$

28
29 **Function** GetNgrams$(labeled\_posts, C)$**:**
30    $ngrams \leftarrow \{\}$
31   **for** *a in all classes* **do**
32   // Get the n-grams of all comments belonging to a post labeled as a
33   **for** *p in labeled_posts[a]* **do**
34   **for** *c in C belonging to post p* **do**
35   Add 1-grams, 2-grams, and 3-grams of comment c to $ngrams[a]$
36   **return** $ngrams$

Figure G5: Subplots of plotting the multi-dimensional scaling from Figure G3 for each pairwise comparison of the 7 classes. The rows and columns follow in the order adoring, amused, approving, excited, angered, saddened, and scared. The entire grid is symmetric for ease of exploration.

Table J7: Examples of posts labeled according to human annotators, CARE, and CARE-BERT. The first three show examples where all three labeling schemes agree, the second three demonstrates examples where external knowledge may be needed, and the last three shows examples where the trajectory of the discussion may be more unpredictable. **Note: CARE-BERT does not get access to the comments.**

Post	Comments	Human	CARE	CARE-BERT
Anxiety: I just want to say that I'm trying...I may not be successful, but I'm trying.	Dude, very proud of you my friend. Don't give up.; Good for you. I'm proud of you for trying. Keep at it.; Happy for you.	approving	approving	approving
AskReddit: What's something you've been wanting to get off your chest but are too scared to?	I'm scared to end up alone and unloved.; I'm so scared to graduate college.; I just got engaged, but I'm not actually happy about it.	saddened; scared	saddened; scared	saddened; scared
AskReddit: What movie really emotionally impacted you?	A Walk to Remember. So sad.; It's a Wonderful Life. Makes me so teary-eyed.; Dead Ringer. Made me so depressed.	saddened; approving	saddened; approving	saddened; approving
Hockey: The Vancouver Canucks have landed a spot in the playoffs!	This is excellent news!; Holy shit, this is exciting!; Hell yeah, fuck the kings!	angered; excited; approving	excited	excited
Panthers: Divisional Playoffs - Panthers vs. 49ers - Discussion Thread Let's do this!	I'M SO MAD; I'm freakin' scared, man.; Screw the whiners! They're going to regret the day they stepped on our turf!	angered; approving	angered	angered; excited
InfertilityBabies: Going to be a line jumper! The doctor says my BP isn't stellar so I am in L and amp;D until Monday morning (37 weeks) induction!	Good luck! So exciting!; Congrats! You're about to be a mom! I'm very excited for you!!!; Super exciting!	excited	excited	approving
AskReddit: What is your favorite TV series ever?	Arrow. It's amazing!; Walking Dead. So excited for the new season!; Teen Titans. It's the best show ever.	approving; excited	approving	approving; amused
Hearthstone: Who is this LIRIK guy, and why does he have 50K subscribers?	This is hilarious; What an idiot. Do more research before posting; He's an adorable guy.	amused; angered; adoring	amused	excited
AskReddit: Imagine that the last thing you ate has been made illegal. What would that be?	Pizza, and now I'm super sad.; Frozen lasagna. Good riddance.; French onion dip. I love that stuff.	approving; saddened	saddened	approving