CR4-NarrEmote: An Open Vocabulary Dataset of Narrative Emotions Derived Using Citizen Science

Andrew Piper
McGill University

Robert Budac
University of Alberta

Abstract

We introduce "Citizen Readers for Narrative Emotions" (*CR4-NarrEmote*), a large-scale, open-vocabulary dataset of narrative emotions derived through a citizen science initiative. Over a four-month period, 3,738 volunteers contributed more than 200,000 emotion annotations across 43,000 passages from long-form fiction and non-fiction, spanning 150 years, twelve genres, and multiple Anglophone cultural contexts. To facilitate model training and comparability, we provide mappings to both dimensional (Valence-Arousal-Dominance) and categorical (NRC Emotion) frameworks. We evaluate annotation reliability using lexical, categorical, and semantic agreement measures, and find substantial alignment between citizen science annotations and expert-generated labels. As the first open-vocabulary resource focused on narrative emotions at scale, *CR4-NarrEmote* provides an important foundation for affective computing and narrative understanding.

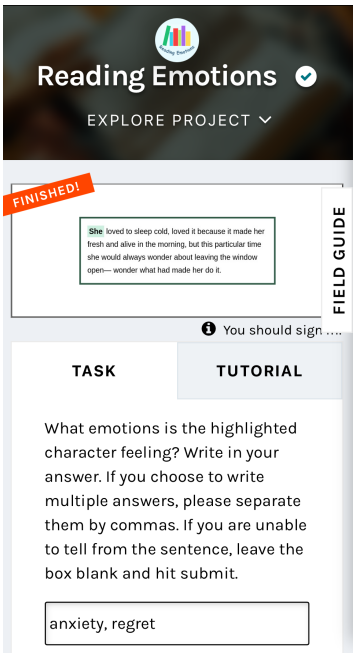

Figure 1: The Citizen Readers project task.

1 Introduction

In this paper, we introduce a new dataset of emotion labels collected through a large-scale citizen science initiative called "Citizen Readers."[1] Over a four-month period, 3,738 volunteers contributed 207,721 annotations of individual sentences drawn from a diverse collection of book-length narratives. Departing from conventional emotion datasets that rely on a fixed set of categories, our annotation paradigm embraces an open vocabulary approach (Wu et al., 2024), inviting annotators to freely describe the emotions they perceived in narrative characters. This open-ended methodology yields a large, fine-grained dataset that captures the nuanced emotional dynamics of human storytelling.

Recent advances in NLP have produced a diverse array of emotion-annotated datasets (Del Arco et al., 2024), spanning domains such as social media (Demszky et al., 2020), dialogue (Li et al., 2017; Rashkin et al., 2019), news headlines (Oberländer et al., 2020), multimodal conversations and screenplays (Tripathi et al., 2018; Poria et al., 2019; Lian et al., 2023), and personal stories (Muhammad et al., 2025). While these resources have enabled significant progress in modeling affective language, they typically rely on fixed emotion taxonomies. Such closed-label approaches, though efficient, limit the expressive range of emotion annotation and tend to suppress culturally or contextually specific emotional states. Open vocabulary methods offer a powerful alternative by allowing emotion labels to emerge from data rather than be predefined (Wu et al., 2024; Lian et al., 2023).

Research in cognitive psychology has shown that narrative reading elicits deep forms of emotional involvement, often driven by perspective-taking and simulated social experience on the part of char-

[1]The full dataset: https://doi.org/10.5683/SP3/XN4ZYZ

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 9785–9796

acters (Koopman, 2015; Mar et al., 2011; Oatley, 2002). Narratologists distinguish such character-centered "narrative emotions," which is our focus here, from more reader- or viewer-centered "aesthetic emotions" (Hogan et al., 2022; Menninghaus et al., 2017). Despite their cultural importance, narrative emotions remain underrepresented in NLP datasets and models, particularly in forms that allow for expressive variability.

Our work contributes to the field of affective computing in three distinct ways:

First, we offer a large number of sentence-level annotations of character emotions **drawn from a diverse array of book-length narrative contexts** (N=43,713). In distinction from recent emotion annotation of personal stories on social media (Muhammad et al., 2025), we focus on long-form stories from twelve different fiction and non-fiction genres, passages spread across the past 150 years, and those drawn from diverse English-language cultural locations including Nigeria, India, and S. Africa in addition to North America.

Second, we provide **free-text emotional responses** by participants, offering a much richer and broader vocabulary of emotional experience than the usual large-scale emotion frameworks commonly in use. As recent research has highlighted, predefined label sets, while useful for standardization, can limit the expressivity and granularity of emotional data (Buechel and Hahn, 2017; Del Arco et al., 2024). By allowing annotators to articulate emotions in their own words, our dataset aligns with recent advances in open vocabulary learning that emphasize flexible, data-driven label discovery beyond pre-defined categories (Wu et al., 2024; Lian et al., 2023).

Third, we demonstrate **the utility of Citizen Science as a viable framework for linguistic annotation**. Research shows that data produced by citizen science projects is of high quality and correlates strongly with expert opinion (McKinley et al., 2017; Kosmala et al., 2016; Wiggins and He, 2016). It also provides a cost-efficient means of data collection (Sauermann and Franzoni, 2015) that addresses concerns surrounding labour ethics (Harmon and Silberman, 2019; Hara et al., 2018) and data quality (Veselovsky et al., 2023; Lu et al., 2020) associated with large-scale crowd-sourcing approaches. Our initiative represents the first of its kind to use citizen science for large-scale story annotation.

In the sections that follow, we describe the data collection process and summary statistics about our data; experiments in mapping granular annotations to continuous VAD variables and discrete emotion categories; manual validation of citizen science labels; insights into the diversity of citizen labels; and finally, benchmarking results using supervised, retrieval-based, and generative models to assess the predictive utility of our dataset under a unified open-vocabulary evaluation framework.

2 Dataset Description

Category	Value
Total annotators	3,738
Total annotations	207,721
1-time annotators	324 (9%)
Heavy contributors (responsible for 80% labels)	656 (18%)
Total passages	43,713
Passages with no character	485 (1%)
Passages with at least 1 label	38,209 (87%)
Passages with multiple labels	30,731 (80%)
Median / Mean emotions per passage	3 / 3.4
Total emotion labels	130,331
Unique emotion labels	1,880
Labels differed from passage	110,569 (85%)
Of all label occurrences:	
Unique labels covering 80%	202 (11%)
Unique labels covering 90%	379 (20%)

Table 1: Summary statistics of the dataset.

2.1 Data Collection Procedure

The data collection process utilized the citizen science platform Zooniverse.org under a project titled "Reading Emotions"[2] and took four months to complete. The project is part of a larger initiative called "Citizen Readers" that aims to enlist the public to help build more transparent, human-centered AI models for understanding human storytelling.

Participants were provided a tutorial, had access to a field guide for more information, a discussion board to post questions, and an "About" page to frame project goals. We had four moderators who would respond daily to user queries. The task was structured in two parts. As can be seen in Figure 1, participants were presented with a single sentence with a highlighted character. They were first asked,

[2] https://www.zooniverse.org/projects/citizenreaders/reading-emotions

"Is the highlighted word a character"? If no, they moved on to the next passage. If yes, they were then asked, "What emotions is the highlighted character feeling? Write in your answer. If you choose to write multiple answers, please separate them by commas. If you are unable to tell from the sentence, leave the box blank and hit submit." We required each sentence to be annotated by at least five different annotators before being retired.

As shown in Table 1, over the four month period, 3,738 participants joined our project and generated a total of 207,721 annotations related to 43,713 unique passages. 9% of participants did one annotation and left the project. 18% of participants completed 80% of annotations. 3,752 passages (ca. 9%) were identified as not having an identifiable emotion by any participants. After cleaning (see A.1), 30,731 passages (ca. 80%) were labeled with more than one emotion by at least one annotator (median = 3, mean = 3.4). We identified a total of 1,880 unique emotion labels with a core set of 202 (11%) accounting for 80% of all occurrences. Finally, we found that only 15% of labels (19,762) matched stems of at least one word from the target sentence suggesting that a vast majority of labels were novel relative to the passage content.

2.2 Data Preparation

Our passages were sourced from three existing datasets: contemporary books spanning twelve different fiction and non-fiction narrative genres (Piper, 2022); fiction books published in the twentieth century (Textual-Optics-Lab, 2025); and works of contemporary fiction published in three non-Western anglophone countries and sampled according to matching criteria (Nigeria, South Africa, and India) (Piper et al., 2025). Table 5 in the Appendix lists the distribution of passages by collection.

All data was first pre-processed using bookNLP (Bamman, 2025). Sentences that were sampled were required to: have a character in the subject position of the sentence who was not a plural entity (entity tag = PER, dependency tag = nsubj, fine-POS-tag does not contain plural); at least one word labeled as "verb.emotion" or "noun.feeling" by BookNLP's super-sense tagger; not be located within dialogue; and belong to a book with a main character that occurred more than twice (i.e. where the character tagging was successful). From these conditions, we then sampled twenty random sentences per book for the Contemporary and Worldlit collections and 2 sentences per book for the 20th-century collection.

2.3 Mapping VAD values and Categorical Emotion Labels

One of the core challenges in working with open-response emotion annotations is the lack of consistency and standardization in the vocabulary used by annotators. While researchers can work directly with citizen labels, we also translate these open-vocabulary labels into two widely-used emotion representation frameworks: (1) dimensional models of affect (Valence, Arousal, Dominance, or VAD) and (2) categorical models based on discrete emotion labels.

2.3.1 VAD Mapping

For VAD, we implement three separate models. First, we directly map each citizen-generated label to the NRC VAD lexicon (Mohammad, 2018), which assigns VAD values to over 20,000 English words. While this approach has the value of consistency, it lacks contextual representation (e.g. all forms of "love" are treated equally).

For our second model, we implement a supervised regression approach using Sentence-BERT (SBERT) embeddings. Rather than relying on direct similarity to predefined VAD scores, we trained a neural network to learn the mapping from word-level SBERT embeddings to continuous Valence, Arousal, and Dominance (VAD) scores. Using the NRC VAD lexicon as training data, we encoded each word into a 384-dimensional SBERT vector and fit a lightweight multi-layer perceptron (MLP) with a hidden layer and sigmoid output activation to predict the three normalized VAD dimensions. Once trained, this model was used to infer VAD scores for our full dataset by feeding it joint embeddings of the citizen-provided label and its narrative context. This approach allows for more context-sensitive emotion modeling that still preserves the weight of the human-labeled emotion.

Third, we trained a fully automated model using the EmoBank corpus (Buechel and Hahn, 2017) to predict VAD scores from sentence embeddings alone, entirely independent of any human-supplied emotion labels. We encoded each sentence using Sentence-BERT and trained the same neural regression architecture as above to predict VAD dimensions directly from the passage-level embeddings. This model was then applied to our dataset to generate purely computational estimates of the affective content of each sentence.

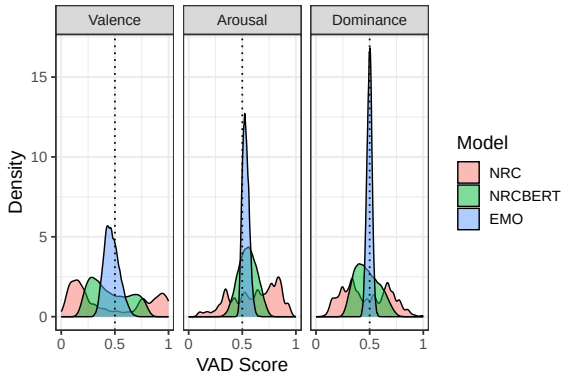

Figure 2: Distributional comparison of Valence, Arousal, and Dominance (VAD) scores across our three mapping models.

As shown in Figure 2, the three VAD mapping models exhibit systematically different distributional characteristics across all dimensions. The NRC lexicon-based model produces the most extreme and bimodal distributions. The NRCBERT model, while similar in shape, shows reduced spread, reflecting the moderating influence of contextualized embeddings on label interpretation. In contrast, the EMO model produces sharply peaked distributions clustered around the neutral midpoint (0.5) for all three dimensions, suggesting that sentence-level embedding models trained on generic corpora tend to underrepresent affective content of narratives.

To evaluate the relative quality of VAD estimates, we conducted a human validation study in which our four expert moderators from the project ranked the outputs of the three models across 200 sample passages. While inter-annotator agreement was moderate (α = 0.31-0.37), aggregate rankings showed consistent patterns. Using binomial tests against a uniform null distribution, NRC was significantly preferred for arousal ($p < 1e\text{-}7$) and dominance ($p < 1e\text{-}6$), and not significantly dispreferred in either, making it the strongest overall choice for these dimensions.

For valence, NRC was both significantly preferred ($p = .003$) and significantly dispreferred ($p = .0006$). Follow-up analysis revealed that NRC was more likely to be disfavored when its valence scores were high ($p < .001$), indicating a possible overestimation of positive emotions. NRCBERT, by contrast, was never significantly dispreferred in any dimension and showed stable performance throughout. We therefore recommend NRC for arousal and dominance tasks, and NRCBERT as the most robust model for balanced valence prediction.

2.3.2 Categorical Emotion Labels

To derive discrete emotion categories, we rely on a similar combined lexicon plus embedding approach. In the first step, we map each citizen label to one or more of the eight basic emotions defined by the NRC Emotion Lexicon: anger, fear, anticipation, trust, surprise, sadness, joy, and disgust (Mohammad and Turney, 2013). Where a label has no direct match, we use the same SBERT embedding method as in the VAD step to identify the nearest label from the NRC emotion vocabulary that also aligns with the sentence's predicted valence. This ensures that predicted emotion labels remain affectively appropriate.

When multiple emotional labels are returned for a given instance, we also select the nearest neighbor in embedding space to the label plus sentence context to produce a single best-fit emotion. Finally, we conduct a valence-emotion consistency check: if a label such as "disgust" is matched to a sentence with positive valence, we adjust the label by substituting the nearest emotion in the embedding space that aligns with the overall valence of the sentence. We find that there are numerous instances in which the lexicon mapping of label to emotion produces directionally inappropriate emotion labels. All original plus adjusted values are available in the released dataset.

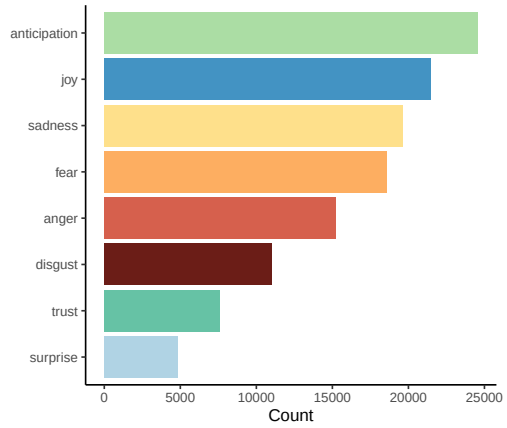

Figure 3: Frequency distribution of NRC emotion categories across all annotated labels.

3 Inter-Rater Reliability of Citizen Science Labels

Assessing inter-rater reliability in open vocabulary emotion annotation presents unique challenges, as

traditional agreement metrics (e.g., Cohen's κ or Krippendorff's α) assume a fixed label set. In our study, annotators freely generated emotion labels, often introducing lexical variation, morphological divergence, and semantic nuance.

We operationalize inter-rater agreement using both exact and approximate matching strategies. First, we compute **exact label overlap** after normalization by measuring the proportion of shared labels between annotators per sentence. Second, we evaluate **category-level agreement** by mapping each label to one of eight NRC-defined basic emotions and calculating the degree to which annotators converge on the same discrete categories. Third, we assess **semantic alignment** by computing the average pairwise cosine similarity between the valence, arousal, and dominance (VAD) values associated with each annotator's label. This captures the degree to which annotators selected affectively similar labels, regardless of lexical or categorical overlap.

3.1 Lexical Overlap

For quantitative assessment, we calculate the average pairwise Jaccard score per sentence, measuring the overlap in emotion labels assigned by different annotators. Across 30,731 passages, we find a mean Jaccard score of 0.118 (SD = 0.173), indicating modest overlap in the exact vocabulary used by different annotators. The median Jaccard score is 0, reflecting the sparsity of exact matches in open-response settings. These results highlight the expressive variability of citizen-generated annotations.

3.2 Categorical Overlap

To complement our string-level analysis, we assess inter-rater agreement at the level of our eight NRC discrete emotion categories. For each passage, we compute the proportion of annotator pairs who selected the same NRC category. Across 29,434 passages with at least two mappable labels, we observe a mean agreement of 0.465 and a median of 0.333, indicating that annotators converge on the same basic emotion category in nearly half of all pairwise comparisons, well above a random baseline of 0.145 (SD = 0.0013). Notably, one-quarter of passages exhibit perfect agreement across all annotators. These results suggest that while annotators frequently differ in lexical choice, they often align in the underlying emotional type being expressed, supporting the interpretive consistency of the open-labeling task.

3.3 Semantic Overlap

To assess whether annotators converged not only lexically or categorically but also semantically, we computed the average pairwise cosine similarity between the valence, arousal, and dominance (VAD) vectors associated with each annotator's label, using the directly mapped NRC VAD lexicon scores for each label. Across all passages with at least two annotations, we found a high mean semantic agreement score of 0.928, indicating strong convergence in affective meaning despite lexical variation. To assess whether semantic convergence exceeded chance, we ran a permutation test that randomly reassigned VAD vectors across annotators while preserving passage structure. Across 1,000 permutations, the mean agreement under the null distribution was markedly lower (M = 0.868, SD = 0.0004), yielding a one-sided p-value < 0.001. This confirms that annotators' labels were not only affectively consistent but significantly more semantically aligned than expected by chance.

4 Expert validation with project moderators

To validate the reliability of our citizen science approach, we conducted an expert validation study using a random sample of 100 passages drawn from our dataset that had been annotated by at least four citizen scientists. These passages were then independently annotated by the four trained moderators of the project using the same interface and open-response format. The four moderators are undergraduate students at McGill who participated in the project design, testing, and moderation for this and prior Citizen Reader projects. This yielded 1,193 total emotion labels in our validation set after cleaning. Given the inherent subjectivity of individual emotion labels, we compare annotations at the dimensional and categorical level using our VAD and discrete emotion mapping described above.

We employed linear mixed-effects models to compare expert assessments with citizen scientist annotations across three emotional dimensions—valence, arousal, and dominance. We use the direct mapping approach for VAD scores in order to condition only on label similarities and not their within-sentence contextualization which will smooth over potential differences. Each model

included a fixed effect for annotator type (moderator vs. citizen scientist) and random intercepts to account for repeated measurements and moderator differences.

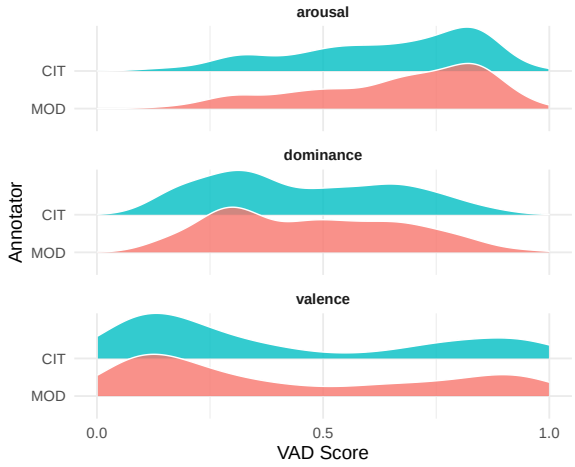

Figure 4: Comparison of the distribution of predicted Valence, Arousal, and Dominance (VAD) scores between citizen annotators (CIT) and project moderators (MOD).

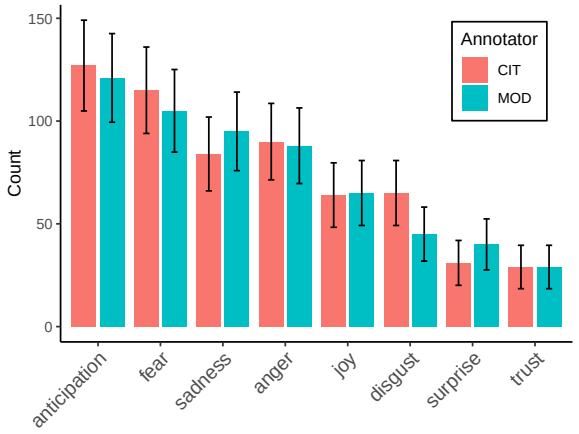

Figure 5: Comparison of citizen annotators (CIT) and moderators (MOD) across discrete emotions.

As can be seen in Figure 4, results showed no statistically significant differences between moderator and citizen annotations for valence ($\beta = -0.016$, $p = .22$), arousal ($\beta = -0.018$, $p = .38$), and dominance ($\beta = -0.021$, $p = .37$). These results indicate that moderator and citizen annotators provide highly consistent affective judgments at the dimensional level.

Further validation mapping individual labels to our eight discrete emotion categories revealed strong consistency between citizen scientists and moderators across all classes (Figure 5). A Poisson regression model predicting annotation count as a function of annotator type, emotion category, and their interaction revealed no significant overall difference in labeling behavior between experts and citizen scientists ($p = 0.881$), and no significant interaction effects between annotator type and any specific emotion category (all $p > 0.05$).

5 Measuring the diversity of citizen scientist labels

One of the values of open-vocabulary labeling approaches is the potential diversity of labels. As we can see in Table 2, while a few of the most common labels belong to the major emotion categories (happy, sad, anger, love) we also find novel emotions that bear on the uniqueness of storytelling, with highly temporally-dependent emotions such as *worried, anxiety, surprised*, and *excited*.

Label	Emotion	Count
hopeful	anticipation	3540
happy	joy	3125
sad	sadness	2901
anger	anger	2871
fear	fear	2664
love	joy	2328
worried	sadness	2229
anxiety	anticipation	1978
surprised	surprise	1768
excited	anticipation	1647

Table 2: Top 10 most frequent emotion labels and their associated NRC emotion categories.

In total, the dataset contains 1,880 unique emotion labels contributed by citizen scientists, exhibiting a pronounced long-tail distribution. As shown in Table 1, the most frequent 202 labels account for 80% of all annotations, while fewer than 400 cover 90%. This distribution indicates a compact yet expressive core vocabulary, supplemented by a diverse set of lower-frequency terms that reflect the heterogeneous ways in which readers interpret emotional content.

To assess the coherence and diversity of our NRC emotion-categorization process, we computed semantic embeddings for these top 202 most frequently used labels using 300-dimensional GloVe vectors (WikiGiga-6b) (Pennington et al., 2014). We included the eight primary emotion categories (e.g., joy, fear, disgust) as anchor terms for

clustering. For each category, we identified the 10 nearest labels in semantic space based on cosine similarity to the anchor emotion category. We then projected each cluster into two dimensions using multidimensional scaling (MDS) (Figure 6).

6 Benchmarking Models Against Citizen Science Labels

To assess the predictive utility of our open vocabulary emotion dataset and establish baseline performance for future modeling, we implement a multipart benchmarking framework that spans supervised, retrieval-based, and generative approaches. Each benchmark is designed to test a distinct inference paradigm—ranging from closed-set classification to open-vocabulary label generation—under consistent evaluation criteria. Inspired by recent work in open vocabulary emotion recognition (Lian et al., 2023), we evaluate models using both conventional classification metrics and set-based overlap measures that better capture the diversity of human judgments.

6.1 Benchmark 1: Supervised Discrete Emotion Classification

In our first benchmark, we establish a closed-set baseline using a supervised classifier trained to predict discrete emotion categories. Each sentence in the dataset is embedded using SBERT (`all-MiniLM-L6-v2`) and a logistic regression classifier is trained to predict the most frequent mapped NRC emotion label per sentence. We use the eight-category NRC framework to enable comparability with prior datasets.

On our dataset, the SBERT + logistic regression model achieves an accuracy of 56.7% and a macro F1 score of 0.53, with the highest performance on *joy* (F1 = 0.68) and the lowest on *trust* and *surprise* (F1 = 0.41 and 0.42, respectively), which were the least represented in our data (Table 3). For comparison, we apply the same model to the GoEmotions dataset, restricted to the six overlapping NRC categories, where it achieves substantially higher performance (accuracy = 72.9%, macro F1 = 0.71). This performance gap indicates greater subtlety and interpretive variability of surrounding narrative emotions.

Emotion	**NarrEmote** (F1)	**GoEmotions** (F1)
Anger	0.56	0.75
Anticipation	0.60	–
Disgust	0.48	0.61
Fear	0.56	0.64
Joy	0.68	0.81
Sadness	0.53	0.69
Surprise	0.42	0.75
Trust	0.41	–
Macro Avg	**0.53**	**0.71**

Table 3: Per-class F1-scores for SBERT + Logistic Regression classifier.

6.2 Benchmark 2: Embedding-Based Semantic Retrieval

Our second benchmark evaluates a zero-shot semantic retrieval approach, in which sentence-level embeddings are compared to a candidate set of emotion label embeddings using cosine similarity. Following the method proposed in Lian et al. (2023), each narrative sentence is embedded using SBERT (`all-MiniLM-L6-v2`), and its top five nearest emotion labels are retrieved from a set of the 202 most frequent citizen-annotated emotion terms, which account for 80% of all label occurrences. No training is involved; similarity is computed directly between sentence and label embeddings.

To evaluate the quality of these top-k predictions, we apply EMER-style set-based overlap metrics. Let $\hat{Y}$ denote the set of predicted emotion labels for a given sentence and Y the set of gold labels aggregated across annotators. We compute *set accuracy* as $\text{Accuracy}_s = \frac{|\hat{Y} \cap Y|}{|\hat{Y}|}$, the proportion of predicted labels that match the gold set, and *set recall* as $\text{Recall}_s = \frac{|\hat{Y} \cap Y|}{|Y|}$, the proportion of gold labels recovered by the model.

As shown in Table 4, the semantic retrieval model achieves scroes well below EMER's English text benchmarks (0.20–0.28) and far from its multimodal results (0.79). This gap likely stems from a combination of annotator freedom and the single-sentence context, where narrative emotions are often implicit and harder to infer.

Model	**Accuracy$_s$**	**Recall$_s$**	**Average**
SBERT	0.0912	0.1787	0.1349
GPT-4o	0.1812	0.1465	0.1639

Table 4: Performance of the evaluation of two open vocabulary emotion inference models on the CR4-NarrEmote dataset.

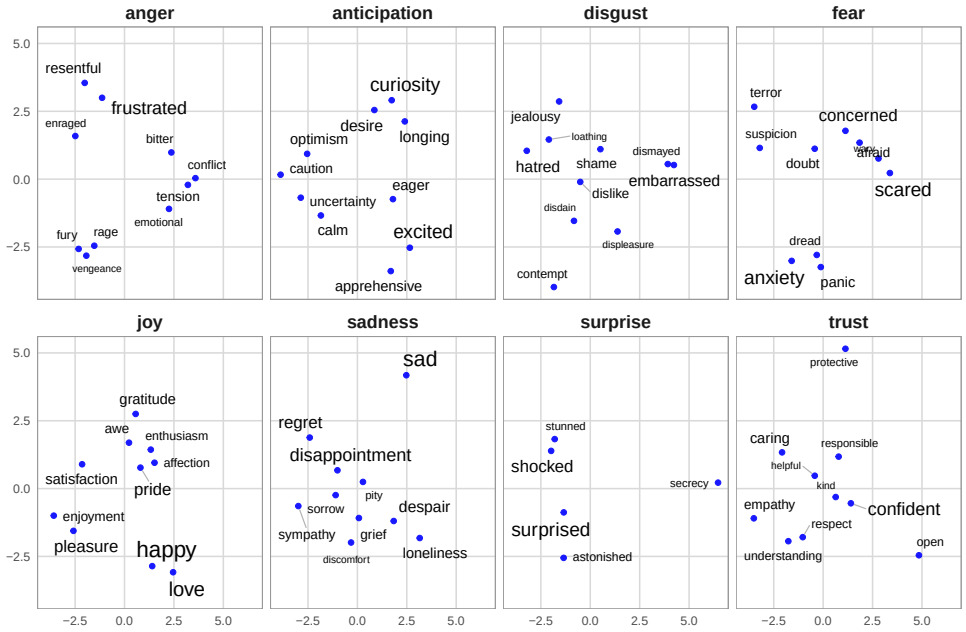

Figure 6: Two-dimensional MDS projection of the top 202 most frequent citizen-supplied emotion labels clustered by proximity to NRC anchor categories using 300-dimensional GloVe embeddings (WikiGiga-6b) of the label and its context. Words are sized by log(frequency).

6.3 Benchmark 3: Prompted Large Language Model Emotion Inference

Our final benchmark tests whether large language models (LLMs) can infer emotion labels from narrative text via direct prompting. For each sentence, we prompt the model with a standardized query that includes the highlighted character and asks for a list of emotions the character is feeling. To ensure structured responses, the model is instructed to return a comma-separated list of single-word emotion labels. For our initial purposes, we evaluate GPT-4o in a zero-shot setting without any fine-tuning or additional context on a sample of 1,000 sentences for illustration purposes.

Predictions are compared against the full set of citizen-provided labels for each sentence using the same EMER-style set-based metrics as in Benchmark 2. As shown in Table 4, GPT-4o is able to recover a meaningful portion of the emotional signal encoded by citizen annotators, despite not being trained on the task or label set. It scores closer to the text-based EMER baselines but still far from multi-modal baselines.

7 Conclusion

This paper introduces "Citizen Readers for Narrative Emotions" (*CR4-NarrEmote*), a large-scale, open-vocabulary dataset of narrative emotions produced through a citizen science initiative. Designed to address a key gap in affective NLP, *CR4-NarrEmote* captures how emotions are expressed and interpreted within narrative settings. Through an open-response annotation format, the dataset surfaces a core vocabulary of ~200 unique emotion terms, offering a rich and nuanced view of human affect in long-form narrative texts.

We map citizen-generated emotion annotations to established affective frameworks—including VAD and NRC emotion categories—using a combination of contextual and lexicon-based models. This alignment enables interoperability with existing affective computing tools while preserving the richness of open-vocabulary responses. Through a combination of lexical, categorical, and semantic agreement measures, we demonstrate that these citizen annotations are both expressive and reliable, closely matching expert judgments.

Benchmarking results show that narrative emotion inference remains a challenging task for current models, with both supervised and generative approaches struggling to match the diversity and subtlety of human annotations. These findings underscore the need for more context-sensitive, narrative-aware emotion models. As an open, extensible resource, *CR4-NarrEmote* lays the groundwork for future research in narrative understanding and affective modeling.

Limitations

While *CR4-NarrEmote* provides a novel resource for studying narrative emotions, it also introduces several limitations that should guide future development. First, the dataset is limited to English-language texts. Although we sample widely across Anglophone regions—including Nigeria, India, and South Africa—the emotional expressions, narrative conventions, and reader inferences reflected here remain culturally and linguistically bounded. As in recent work (Muhammad et al., 2025), expanding to multilingual corpora will be essential for examining how narrative emotions vary across different storytelling traditions.

Second, the dataset captures emotions at the sentence level, which is a pragmatic but reductive unit of narrative analysis. We pre-tested numerous contextual frameworks prior to launching the project and found that single sentences provided the ideal balance between clarity and interpretibility (you get less bang for your contextual buck the longer the context window). Nevertheless, future work will want to experiment with significantly larger contexts to better understand the duration of emotional states. Similarly, the annotation task isolates single characters per sentence, potentially underrepresenting emotions that emerge from interpersonal dynamics or broader plot contexts. Incorporating paragraph- or character-arc level annotations could improve alignment with how emotions are experienced and interpreted in narrative form.

Third, while open-vocabulary labeling supports expressive richness, it introduces challenges in consistency and model training. Despite rigorous cleaning and mapping pipelines, annotation variability and lexical sparsity in the long tail of emotion terms complicate both inter-rater agreement and benchmarking. Moreover, while we benchmark several inference models—including contextual embeddings and LLMs—the results reflect the difficulty of recovering subtle narrative inferences from isolated text. Many citizen-supplied labels reflect interpretive leaps or emotional resonance not explicitly stated in the sentence, posing a significant challenge for existing affective computing frameworks.

Ethical Considerations

This research was conducted with the approval of our institutional Research Ethics Board (REB File number: 22-04-076), ensuring that all procedures met ethical standards for human participant research. Participation in the "Citizen Readers" project was entirely voluntary, anonymous, and conducted through the Zooniverse.org platform, which provides public-facing information on project goals, data use, and moderation protocols. No identifying information was collected, and participants could withdraw at any time without consequence. A detailed "About" page, annotation tutorial, and discussion board were provided to ensure informed participation.

We recognize the ethical significance of engaging volunteers in data annotation, particularly in light of growing concerns around exploitative crowd-sourcing practices. Our citizen science approach aims to foreground collaborative knowledge production rather than extractive labor. Unlike paid microtask platforms, Zooniverse is structured to support public engagement, transparency, and participant agency. Moderators provided support throughout the project, and care was taken to ensure tasks were cognitively meaningful and non-repetitive. Nonetheless, we acknowledge that not all participants may share the same levels of digital literacy or interpretive experience, and future work could explore more targeted onboarding or feedback mechanisms.

Emotion annotation poses specific ethical challenges. Emotional language is culturally shaped and often subjective, raising questions about interpretive bias, projection, and the limits of consensus. While our open-vocabulary design aims to preserve the expressive nuance of annotator interpretations, it also introduces ambiguity that may complicate downstream use. To mitigate risks of misrepresentation, we provide detailed mappings, confidence metrics, and full transparency around label provenance. We also avoid any diagnostic use of emotion data, emphasizing that the labels reflect perceived character states within fictional narratives, not mental health or real-world affect.

Finally, we believe the open sharing of data and methods contributes to more transparent and equitable AI development. All data, annotation protocols, and evaluation benchmarks are publicly released to support reproducibility, critique, and broader community use. We encourage researchers building upon CR4-NarrEmote to consider the cultural and interpretive complexity of narrative emotions, and to engage with ethical questions not only in annotation practices, but also in model deployment and interpretation.

Acknowledgments

We would like to thank the Social Sciences and Humanities Research Council of Canada (435-2022-0089) for generous funding to support this project.

References

David Bamman. 2025. Booknlp: A natural language processing pipeline for books. https://github.com/booknlp/booknlp. Accessed: 2025-04-01.

Sven Buechel and Udo Hahn. 2017. Emobank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 578–585.

Flor Miriam Plaza Del Arco, Alba A Cercas Curry, Amanda Cercas Curry, and Dirk Hovy. 2024. Emotion analysis in nlp: Trends, gaps and roadmap for future directions. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5696–5710.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*.

Kotaro Hara, Abigail Adams, Kristy Milland, Saiph Savage, Chris Callison-Burch, and Jeffrey P Bigham. 2018. A data-driven analysis of workers' earnings on amazon mechanical turk. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–14.

Ellie Harmon and M Six Silberman. 2019. Rating working conditions on digital labor platforms. *Computer Supported Cooperative Work (CSCW)*, 28(5):911–960.

Patrick Colm Hogan, Bradley J Irish, and Lalita Pandit Hogan. 2022. *The Routledge companion to literature and emotion*. Routledge London.

Eva Maria Emy Koopman. 2015. Empathic reactions after reading: The role of genre, personal factors and affective responses. *Poetics*, 50:62–79.

Margaret Kosmala, Andrea Wiggins, Alexandra Swanson, and Brooke Simmons. 2016. Assessing data quality in citizen science. *Frontiers in Ecology and the Environment*, 14(10):551–560.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995.

Zheng Lian, Licai Sun, Mingyu Xu, Haiyang Sun, Ke Xu, Zhuofan Wen, Shun Chen, Bin Liu, and Jianhua Tao. 2023. Explainable multimodal emotion reasoning. *CoRR*.

Jian Lu, Wei Li, Qingren Wang, and Yiwen Zhang. 2020. Research on data quality control of crowdsourcing annotation: A survey. In *2020 IEEE Intl Conf on dependable, autonomic and secure computing, Intl Conf on pervasive intelligence and computing, Intl Conf on cloud and big data computing, Intl Conf on cyber science and technology congress (DASC/PiCom/CBDCom/CyberSciTech)*, pages 201–208. IEEE.

Raymond A Mar, Keith Oatley, Maja Djikic, and Justin Mullin. 2011. Emotion and narrative fiction: Interactive influences before, during, and after reading. *Cognition & emotion*, 25(5):818–833.

Duncan C McKinley, Abe J Miller-Rushing, Heidi L Ballard, Rick Bonney, Hutch Brown, Susan C Cook-Patton, Daniel M Evans, Rebecca A French, Julia K Parrish, Tina B Phillips, and 1 others. 2017. Citizen science can improve conservation science, natural resource management, and environmental protection. *Biological conservation*, 208:15–28.

Winfried Menninghaus, Valentin Wagner, Julian Hanich, Eugen Wassiliwizky, Thomas Jacobsen, and Stefan Koelsch. 2017. The distancing-embracing model of the enjoyment of negative emotions in art reception. *Behavioral and Brain Sciences*, 40:e347.

Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 174–184.

Saif M Mohammad and Peter D Turney. 2013. Nrc emotion lexicon. *National Research Council, Canada*, 2:234.

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine de Kock, Nirmal Surange, Daniela Teodorescu, Ibrahim Said Ahmad, and 1 others. 2025. Brighter: Bridging the gap in human-annotated textual emotion recognition datasets for 28 languages. *arXiv preprint arXiv:2502.11926*.

Keith Oatley. 2002. Emotions and the story worlds of fiction. *Narrative impact: Social and cognitive foundations*, 39:69.

Laura Ana Maria Oberländer, Evgeny Kim, and Roman Klinger. 2020. Goodnewseveryone: A corpus of news headlines annotated with emotions, semantic roles, and reader perception. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1554–1566.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Andrew Piper. 2022. The conlit dataset of contemporary literature. *Journal of Open Humanities Data*, 8.

Andrew Piper, David Bamman, Christina Han, Jens Bjerring-Hansen, Hoyt Long, Itay Marienberg-Milikowsky, Tom McEnaney, Mathias Iroro Orhero, Emrah Peksoy, Pallavi Rastogi, and 1 others. 2025. Mini worldlit: A dataset of contemporary fiction from 13 countries, nine languages, and five continents. *Journal of Open Humanities Data*, 11(1).

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381.

Henry Sauermann and Chiara Franzoni. 2015. Crowd science user contribution patterns and their implications. *Proceedings of the national academy of sciences*, 112(3):679–684.

Textual-Optics-Lab. 2025. Us novel corpus. Accessed: 2025-03-29.

Samarth Tripathi, Sarthak Tripathi, and Homayoon Beigi. 2018. Multi-modal emotion recognition on iemocap dataset using deep learning. *arXiv preprint arXiv:1804.05788*.

Veniamin Veselovsky, Manoel Horta Ribeiro, and Robert West. 2023. Artificial artificial artificial intelligence: Crowd workers widely use large language models for text production tasks. *arXiv preprint arXiv:2306.07899*.

Andrea Wiggins and Yurong He. 2016. Community-based data validation practices in citizen science. In *Proceedings of the 19th ACM Conference on computer-supported cooperative work & social computing*, pages 1548–1559.

Jianzong Wu, Xiangtai Li, Shilin Xu, Haobo Yuan, Henghui Ding, Yibo Yang, Xia Li, Jiangning Zhang, Yunhai Tong, Xudong Jiang, and 1 others. 2024. Towards open vocabulary learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(7):5092–5113.

A Appendix

A.1 Data Cleaning Procedure

In order to clean our free-text responses, we undertook the following steps. We used the NRC VAD Lexicon as our emotion reference set, which contains 20,007 unique emotion labels. All labels were lowercased and whitespace stripped. We first matched all normalized citizen science labels to the VAD lexicon. This left us with a total of 3,705 unmatched unique labels and 10,669 total labels (out of 132,958 initial rows). Next we ran a standard spellcheck over the remaining rows (proffessional → professional, fasination → fascination, etc.). This resulted in another 1,553 unique label corrections (2,933 total rows). For the remaining unmatched labels, we stemmed both the labels and the VAD lexicon and substituted the VAD full word for every matched stem. This resulted in another 4,955 matches, leaving 1,194 unique unmatched labels and 3,346 rows. We next ran all remaining unique labels through GPT-4o. We asked it to translate any non-English emotion labels and remove any words that could not plausibly be associated with a character's emotions. This resulted in the removal of another 455 unique labels / 694 rows. We then removed any remaining labels that matched a standard stopword list. This left us with 669 unmatched unique labels / 2,517 rows. We then remove all labels that appear only once.

As a next round of cleaning, we engage in two further aggregation steps. Here we stem all labels using the Porter stemmer. Then for all identical stems we aggregate them using the most frequent original label. For example, abandon (4), abandoned (76), and abandonment (12), would all be reconciled to "abandoned" as the most frequent representative of the stem. For remaining non-aggregated but morphologically related words, we supply the remaining 2,222 labels to GPT-o1 in batches asking it to resolve morphologically similar words that are not antonyms. This resulted in 394 aggregations which we then manually reviewed by hand. Successful aggregations missed by the stemmer were identified (anger, angry; puzzlement, puzzled, etc.), but several mistakes needed to be corrected (hopeful, hopeless; tempermental, tempered, etc.). Notably we do not aggregate synonyms such as joy/happiness or compassion/empathy, etc., in order to leave in place the diversity of annotator vocabulary. This gives us our final number dataset of 130,331 annotations and 1,880 unique labels.

A.2 Passage Counts by Collection

Contemporary (FIC)		**Contemporary (NON)**		**Worldlit**	
Bestsellers (BS)	2388	Biographies (BIO)	1833	India	1739
Middle-Grade (MID)	1572	Histories (HIST)	1906	Nigeria	1303
Mysteries (MY)	2238	Memoirs (MEM)	2127	South Africa	1470
NY Times Reviewed (NYT)	4018	Mixed (MIX)	1779		
Prizewinners (PW)	2483				
Science-Fiction (SF)	2154				
Young Adult (YA)	1666				
Total	**16,519**	**Total**	**7,645**	**Total**	**4,512**

Twentieth-Century

Decade	1880	1890	1900	1910	1920	1930	1940
Passages	173	309	594	696	775	1068	1110
Decade	1950	1960	1970	1980	1990	2000	**Total**
Passages	773	810	1384	2574	4046	589	**14,999**

Table 5: Counts of passages across all genres, periods and regions.