My Heart Skipped a Beat!
Recognizing Expressions of Embodied Emotion in Natural Language

Yuan Zhuang[1], **Tianyu Jiang**[2], **Ellen Riloff**[3]
[1]University of Utah, [2]University of Cincinnati, [3]University of Arizona
yuan.zhuang@utah.edu, tianyu.jiang@uc.edu, riloff@cs.arizona.edu

Abstract

Humans frequently experience emotions. When emotions arise, they affect not only our mental state but can also change our physical state. For example, we often open our eyes wide when we are surprised, or clap our hands when we feel excited. Physical manifestations of emotions are referred to as *embodied emotion* in the psychology literature. From an NLP perspective, recognizing descriptions of physical movements or physiological responses associated with emotions is a type of implicit emotion recognition. Our work introduces a new task of recognizing expressions of embodied emotion in natural language. We create a dataset of sentences that contains 7,300 body part mentions with human annotations for embodied emotion. We develop a classification model for this task and present two methods to acquire weakly labeled instances of embodied emotion by extracting emotional manner expressions and by prompting a language model. Our experiments show that the weakly labeled data can train an effective classification model without gold data, and can also improve performance when combined with gold data. Our dataset is publicly available at https://github.com/yyzhuang1991/Embodied-Emotions.

1 Introduction

Most people experience emotions every day. When emotions arise, we not only feel them mentally but we also experience them physically via our body. Sometimes an emotion evokes a visible physical reaction. For instance, we may clench our fists or stomp our feet when we feel angry, or raise our hands in the air and dance when we feel happy. We may also have physiological responses when we experience an emotion. For example, we may feel our heart racing or feel a chill down our spine when we get scared. Or we may feel our cheeks flush when we are embarrassed. In general, the physical experience of an emotion via our body is referred to as **embodied emotion** in the psychology literature (Lakoff and Johnson, 1999; Prinz, 2004; Niedenthal, 2007; Barrett et al., 2008), and it has been recognized as an important component of emotional experiences. Figure 1 shows examples of body part references that are and are not associated with embodied emotions.

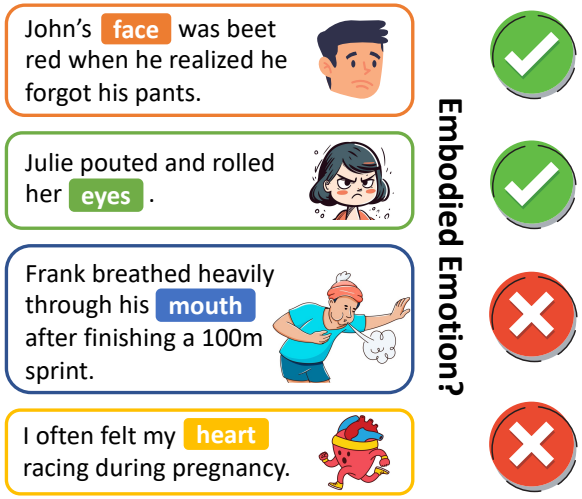

Figure 1: Illustration of body part mentions associated with embodied emotions, or not related to emotion.

Recognizing expressions of embodied emotion in natural language is important to identify implicit emotional states, which is a major challenge in emotion recognition (Alm et al., 2005; Mohammad and Turney, 2010; Mohammad et al., 2018). For example, if we read that *"John slammed his fist against the wall"*, we would infer that John is angry. Similarly, if Jane says *"My hands sweated profusely before my presentation"*, we understand that Jane was nervous. In addition, recognizing embodied emotion expressions could help identify behavioral traits and monitor problematic behaviors such as antisocial behaviors (Parrott, 2001; Munezero et al., 2011), which are closely tied to physical responses stimulated by negative emotions.

Our work introduces the first study on recognizing expressions of embodied emotion in natural language. We formalize the task as a classification problem to determine whether a body part reference describes an embodied emotion. We have created a benchmark dataset, **CHEER**, which contains 7,300 body part mentions with human annotations for this task. We conduct extensive experiments to evaluate the effectiveness of multiple existing emotion classifiers on our dataset and show that they do not perform well at recognizing embodied emotion expressions.

We also present two methods to automatically produce weakly labeled data for this task. We develop a pattern-based method that identifies body part words that are syntactically connected to emotion words through manner expressions. For example, *"He slammed his fist in anger"* reveals that *"slammed his fist"* is an embodied reaction to anger. The second method identifies instances of embodied emotion based on prompting a large language model (LLM). Our experiments show that the resulting weakly labeled data could be used to train an effective classifier and also improve classification performance when combined with gold data. To sum up, our contributions are three-fold:

1. We introduce a novel task of recognizing expressions of embodied emotion in natural language. We create a dataset of 7,300 body part mentions with human annotations indicating whether the body part is involved in an embodied emotion. The dataset can be found at `https://github.com/yyzhuang1991/Embodied-Emotions`.

2. We conduct extensive experiments to evaluate multiple existing emotion classification models on this task and show that they do not perform well.

3. We propose two methods to produce a large set of weakly labeled instances for this task. We show that the weakly labeled data can be used to train an effective embodied emotion classifier and also improve classification performance when combined with the gold data.

2 Related Work

In NLP, emotion recognition has been extensively studied. Researchers have worked on creating emotion resources (Strapparava and Valitutti, 2004; Mohammad and Turney, 2010, 2013; Demszky et al., 2020) and analyzing emotion in texts across different genres (Rosenthal et al., 2017; Mohammad et al., 2018; Demszky et al., 2020). Some research has focused on identifying implicit emotion/affect, including implicit sentiment analysis (Li et al., 2021), good-for/bad-for events (Deng and Wiebe, 2014, 2015) and affective event recognition (Ding and Riloff, 2018; Zhuang et al., 2020; Zhuang and Riloff, 2023). Another relevant line of work is the study of non-verbal communication signals for expressing emotions (Kim and Klinger, 2019), such as physical appearance, facial expressions and movements of body as a whole. The non-verbal communication signals are a superset of embodied emotions, and Kim and Klinger (2019) only performed manual analyses and did not propose an automated task for recognizing these non-verbal signals. Another similar line of work (Casel et al., 2021; Cortal et al., 2023), inspired by the Emotion Component Process Model (Scherer, 2005), focuses on recognizing emotion components that are often used for expressing emotions. Casel et al. (2021) identifies five emotion components, including Cognitive Appraisal, Neurophysiological Symptoms, Motivational Action Tendencies, Motor Expressions and Subjective Feelings. Similarly, Cortal et al. (2023) identifies four emotion components, including Behavior, Feeling, Thinking and Territory. This line of work differs from ours in two aspects. First, the emotion components are fundamentally different from embodied emotions. For example, Cortal et al. (2023) includes all behaviors not evoked by emotion during an emotional event (e.g., *giving a lecture*) and Casel et al. (2021) includes goal-oriented physical movements (e.g., *recover the stolen horse*), while ours does not. In addition, their work focuses on teasing apart different emotion components from each other. In essence, it assumes the text to classify is emotional, while ours does not.

Researchers have found that figurative language is commonly used to express emotion (Ghosh et al., 2015), including metaphor (Mohammad et al., 2016), sarcasm (González-Ibáñez et al., 2011) and rhetorical questions (Zhuang and Riloff, 2020). We observe that embodied emotion expressions are sometimes metaphorical. For example, the phrase "*butterflies in my stomach*" refers to a true physiological reaction to an emotion (so it is an embodied emotion) but the physical sensation is described metaphorically. However, many metaphorical expressions that mention a body part do not corre-

spond to an actual physical response. For example, the phrase "*my heart melted*" is often used emotionally, but it does not indicate a real physical change of the heart.

In psychology, most emotion theories acknowledge that the body plays a role in the emotion experience (Barrett et al., 2008). Different theories have been proposed for the role of the body when emotion arises. Some studies suggest that changes in the body cause emotion (James, 1884; Ekman, 1972). On the other hand, some researchers propose that emotion results in bodily changes (Darwin, 1872; Arnold, 1968; Frijda, 1986). And some modern theories of embodied emotion (Lakoff and Johnson, 1999; Barsalou and Wiemer-Hastings, 2005) present a different view that the mind and the body interact and shape emotion. Our work draws heavily from the view that emotion results in bodily changes.

3 Task and Dataset

3.1 Task Definition

We propose a new task to recognize expressions of *embodied emotion* in natural language. While emotion could be embodied in one body part, multiple body parts, or even the whole body, we focus on recognizing expressions of embodied emotion in a single body part.

We formulate the task as a binary classification problem, which classifies a body part word within some context into one of the following two categories: 1) **Embodied Emotion (EE)**, where the body part is involved in embodied emotion; 2) **Neutral**, where the body part is not involved in embodied emotion. We define the task as follows:

Definition: *A body part is involved in an embodied emotion if both conditions below are satisfied*:

1) *A physical movement or physiological arousal involving the body part is evoked by emotion.*

2) *The physical movement, if there is any, has no purpose other than emotion expression.*

Condition 1 requires that the physical reaction is caused by emotion. This excludes reactions from other causes, such as weak legs after exercising or watery eyes because of allergies. Condition 2 applies to physical movements (not physiological arousals) and requires that the physical movement has no other purpose. This condition excludes movements that also aim to accomplish a goal. For example, consider the scenario where a house fly is annoying someone, so they slam it with their fist. This action is motivated by emotion, but it is also intended to kill the fly. The set of actions that could be motivated by an emotion are nearly limitless, and the degree to which an emotion causes an action is often ambiguous. Our definition of embodied emotion focuses on movements and physiological arousals that are *solely* emotional and have no additional goal.

Our task is also contextualized. We identify embodied emotion based on a sentence and its preceding context because physical reactions can be ambiguous without context. For example, the phrase *"my heart is racing"* is likely an expression of embodied emotion in the context of a scary situation, but not in the context of physical exercise.

3.2 Data Collection

Our first goal was to build a dataset of sentences that mention body part words. We began by collecting the terms in two online word lists of body part vocabulary.[1] Then we filtered the list by removing multi-word phrases (e.g., "index finger") and plurals. We removed multi-word phrases because most of those phrases in the list referred to internal organs that are rarely discussed and unlikely to be associated with emotions (e.g., "lumbar vertebrae"). After the filtering step, the final list contains 162 body part words.

Next, we extracted sentences that mention these body parts in the personal blogs that Ding and Riloff (2018) extracted from the ICWSM 2009 and 2011 Spinn3r datasets (Kevin et al., 2009, 2011). This resulted in around 3 million sentences. It is often insufficient to identify embodied emotion based on one sentence in isolation, so we also kept the three preceding sentences.

We next performed several preprocessing steps to clean the collected texts. We used CoreNLP (Manning et al., 2014) to facilitate this process, such as tokenization and named entity recognition. First, we observed that the data included a lot of sexual descriptions. Sexuality and emotions are often intertwined and determining whether physical responses related to sexual encounters are truly evoked by emotion is challenging, so we decided to exclude texts with sexual descriptions. Specifically, we discarded sentences

[1] https://www.collinsdictionary.com/us/word-lists/body-parts-of-the-body and https://www.enchantedlearning.com/wordlist/body.shtml

that contain words in the Sexual category of the LIWC lexicon (Pennebaker et al., 1999). We also excluded body part mentions (i.e., did not label them) that occur in contexts that mention multiple people because they are also frequently romantic situations. Specifically, we excluded a body part mention if the 5-word window around it contains a plural personal pronoun or at least two different person mentions (personal pronouns or named entities).[2] Finally, we ignored body part mentions that are preceded by a second-person possessive pronoun or a third-person possessive (not pronoun) because these usually refer to another person ("your eyes") or a non-human entity (e.g., "the cat's head"). We leave for future work the challenge of disentangling emotions and physical actions in multi-person event descriptions.

Finally, we removed infrequent body parts because they usually refer to very specific body parts that are rarely associated with emotions (e.g., "*epiglottis*" and "*ulna*"). We excluded body parts that occurred in less than 0.1% of the sentences. This process produced a final dataset of 868,003 sentences with 56 distinct body parts.

3.3 Gold Standard Annotation

We asked two people to produce the gold annotations. An annotation instance is a body part mention in a sentence and the three preceding sentences as context. The annotators produced a binary label (Embodied Emotion vs. Neutral) to indicate if the body part is associated with an embodied emotion, following the definition in Section 3.1.

We asked the annotators to annotate 2,600 randomly selected sentences that mention a body part. If a sentence mentioned multiple body parts, each mention was presented as a separate instance to annotate. This process produced 2,948 annotated body part mentions. The pairwise inter-annotator agreement measured by Cohen's Kappa was 0.79, indicating good agreement. The annotators adjudicated their disagreements to produce the final gold labels. We used this data as the test set. We then asked the annotators to individually label more data and randomly split these instances into a training set and validation set by the ratio of 7:3. We also made sure that annotation instances that belong to the same sentence went into the same set.

The complete dataset contains 56 distinct body part mentions and 7,300 annotated instances, which consist of 1,350 (18.5%) Embodied Emotion and 5,950 (81.5%) Neutral. Appendix A contains the full list of body parts and more statistics. We will refer to this dataset as **CHEER** (a **C**ollection of **H**uman annotations for **E**mbodied **E**motion **R**ecognition). Table 1 shows the detailed dataset statistics. Table 2 shows Embodied Emotion instances in the CHEER data.

[2]The 5-window is not applied across sentences.

	EE (%)	Neutral (%)	Total
Train	578 (19.1%)	2,452 (80.9%)	3,030
Validation	264 (20.0%)	1,058 (80.0%)	1,322
Test	508 (17.2 %)	2,440 (82.8%)	2,948
Total	1,350 (18.5%)	5,950 (81.5%)	7,300

Table 1: Statistics of CHEER in terms of annotated body part mentions. **EE**: Embodied Emotion.

- Kiki came to me and jump onto my lap. I rolled my **eyes** and went "Stupid cat".
- He sits down and tells me he is going to need my social, and all my names I've had in my life. Immediately my **throat** tightens.
- When we got home, she'd been brooding and pouting and stomping her **feet** as she sulked around the house with nothing to do.
- She started to shake her head, to deny it all yet again ... A loud sob raced up the back of her throat, choking her, and her **knees** buckled.
- "You let him pay for your meal?" He felt his **eyebrows** fly up in astonishment.

Table 2: Embodied Emotion examples in CHEER. The preceding contexts are shortened for brevity.

4 Evaluating Emotion Classifiers

We conducted experiments to investigate how well existing emotion classifiers recognize embodied emotion. We evaluated several classifiers that achieved state-of-the-art performance on emotion or affect recognition tasks. The implementation details of these classifiers can be found in Appendix B. The first model is **SpanEmo** (Alhuzali and Ananiadou, 2021), which is based on BERT (Devlin et al., 2019) and trained on the affective tweets in SemEval-2018 (Mohammad et al., 2018). The second model, which we will refer to as **GE-BERT**, is a BERT-base model fine-tuned with the GoEmotions data in (Demszky et al., 2020). We also evaluated **Seq2Emo** (Huang et al., 2021), a Bi-LSTM trained on the GoEmotions Dataset. All

these models take a text snippet as input and generate multi-label emotions. Finally, we evaluated **Aff-BERT** (Zhuang et al., 2020), an affective event classifier that takes an event phrase as input and identifies its affective polarity.

These models were trained on different types of input, so we experimented with four strategies for applying each classifier to instances in our CHEER data. Consider the instance below with the underlined "*eyes*" as the targeted body part:

> **Preceding Context**: *Every step he took echoed throughout the room. He stood in front of me, empty eyes locked into mine.*
>
> **Sentence**: *Then my eyes instantly widened and my mouth dropped open.*

The first two strategies provide full sentences as input to a classifier: a) **Multi-sent**: the input is the preceding context concatenated with the sentence that mentions the body part. b) **Sent**: the input is just the sentence that mentions the body part.

The next two strategies zero in on the context immediately surrounding the body part mention: c) **Window**: the input is the k-word window around the body part mention (e.g., the 2-word window is "*Then my eyes instantly widened*"); d) **Event**: the input is the event phrase that mentions the body part (e.g., "*my eyes widened*"). We extract events from dependency parse trees following the same representation used by Aff-BERT. In all cases, the instance is tagged with embodied emotion if the classifier recognizes the corresponding input as emotional/affective.[3]

Method	Macro F1	EE			Neutral		
		Pre	Rec	F1	Pre	Rec	F1
SpanEmo†	45.2	18.4	53.7	27.4	83.9	50.4	63.0
Aff-BERT†	50.3	21.7	**56.1**	**31.3**	**86.4**	57.8	69.3
Seq2Emo*	54.4	**28.8**	16.7	21.2	84.1	**91.4**	**87.6**
GE-BERT*	**58.2**	31.0	30.3	30.6	85.6	85.9	85.7

Table 3: Test performance of emotion classification models. †: the *Event* strategy. *: the *Window* strategy.

Experimental Results Table 3 shows the performance of the emotion classifiers on the test set of CHEER. We tried all 4 strategies for all classifiers (except Aff-BERT which requires *Event* representations) and show the best result for each classifier in Table 3. The full results can be found in Appendix B.1. The *Window* strategy performed best for Seq2Emo and GE-BERT, while the *Event* representation performed best for SpanEmo. Surprisingly, all models performed best without the contextual information.

Overall, GE-BERT produced the best macro F1 score of 58.2% . However, it only achieved about 30% recall and precision for recognizing embodied emotions. These results demonstrate that embodied emotions cannot be reliably recognized by existing methods for emotion recognition, which motivates the need for further research on this topic.

5 Producing Weakly Labeled Data for Embodied Emotions

Our goal was to create a classifier for recognizing embodied emotion expressions. We created gold training data, but its amount is relatively small as human annotation is time-consuming. In this section, we introduce two methods to automatically produce a large amount of weakly labeled instances. We will later show that this weakly labeled data can be used to train an effective classifier without any gold data at all, or used in combination with gold data to further improve classification performance.

5.1 Labeling Data by Dependency Patterns

Our first method produces new Embodied Emotion instances by identifying body part words that are syntactically connected to an explicit emotion word through a manner expression. Specifically, we extract two types of manner expressions:

- Prepositional phrases with "in" or "with" and an emotional head noun (e.g., "*My mouth opened **in surprise***" or "*I clapped my hands **with great excitement***").
- Emotional adverb (e.g., "*I **angrily** clenched my fists*" or "*I **impatiently** tapped my finger*").

We observed that emotional manner expressions in the forms above often describe a physical experience when emotion arises (e.g., "*I angrily broke the window*"). As a result, when a body part is syntactically connected to such emotional manner expressions, the sentence tends to describe the physical experience of emotion via the body part.

For emotional nouns in prepositional phrases, we used all positive and negative nouns with strong subjectivity (641 nouns in total) in the MPQA lexicon (Wilson et al., 2005). For emotional adverbs,

[3] If the body part is mentioned in multiple event phrases, we label the instance as embodied emotion if any of the phrases is tagged as emotional/affective by the classifier.

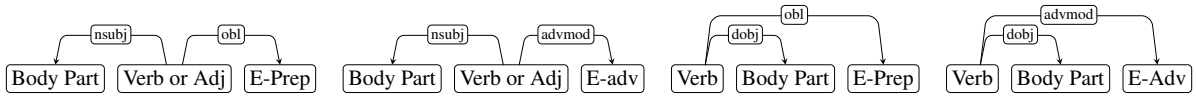

Figure 2: Dependency relation patterns. E-Prep is a prepositional phrase with emotion head noun. E-Adv is an emotion adverb.

Instruction

You will need to determine if a body part is involved in any embodied emotion. Specifically, a body part is involved in some embodied emotion if both conditions below are satisfied: 1) The physical movement or physiological arousal involving the body part is evoked by emotion. 2) The physical movement, if there is any, has no other purpose than emotion expression.

Instance

Input: My heart still flutters when I think about it.

Question: Is the body part "my heart" in Input involved in any embodied emotion? No explanation.

Figure 3: Example prompt for GPT3.5. Input placeholders are wrapped by boxes in red.

we leveraged the WordNet Affect lexicon (Strapparava and Valitutti, 2004), which associates a subset of words in WordNet (Miller, 1992) with 28 emotions. We extracted the 121 adverbs that are associated with the 6 basic Ekman's emotions (Ekman, 1992).

Our pattern-based method first extracts sentences that contains a body part word in one of the emotional manner expressions described earlier. We create an Embodied Emotion instance if a body part word is connected to an emotional manner expression matching one of the dependency relation patterns illustrated in Figure 2. Finally, we remove the emotional manner expression from the sentence so that the classifier cannot use it when learning to recognize embodied emotions.

5.2 Labeling Data by LLM Prompting

The pattern-based method is not able to harvest Neutral Instances. In addition, the diversity of the harvested instances may be limited because some body parts rarely co-occur with the emotional manner expressions. To overcome these issues, we also produced new labeled instances by prompting a large language model (GPT3.5). Specifically, we construct a template with an instruction and input placeholders. Given an input instance, we fill the input placeholders with the body part and the sentence that mentions it (see an example in Figure 3) and feed it to the language model.[4] We then assign the label based on the yes-or-no answer.

[4]Note that the preceding sentences are not used in the prompt. We found that using them hurt performance.

5.3 Weakly Labeled Dataset

We applied both methods to the subset of the 868,003 sentences in Section 3.2 that were not labeled by the annotators. The pattern-based method produced 7,162 Embodied Emotion instances. For the prompting method, we used GPT3.5 because it achieved the best zero-shot performance (see Section 6.1). We first applied the prompting method to collect 7,000 Embodied Emotion instances.[5] Then we continued to collect 56,648 Neutral instances to maintain a distribution of 20% Embodied Emotion and 80% Neutral, given that there are 14,162 Embodied Emotion instances.

6 Experimental Results

We conducted experiments with classification models trained on weakly labeled data, gold labeled data, or both. We also present results for zero-shot prompting with LLMs as a baseline comparison. For the evaluation metric, we report the macro-averaged F1 score over the test set, as well as Precision, Recall and F1 for each class.

Our classification model is based on fine-tuning the pretrained BERT model (Devlin et al., 2019). Given an input instance, we concatenate the preceding sentences and the sentence that mentions the body part, and insert the CLS token between them. We pass this to the BERT-base-uncased model and get its last-layer token embeddings. Finally, we produce an embedding for the body part word by averaging the embeddings of its leftmost and rightmost tokens, and then feed it through a linear classification layer to predict the label. For the sake of brevity, we will refer to the classification model as the Embodied Emotion Classifier (**EEC**).

6.1 Baselines & Gold Supervision

Large language models (LLMs) have shown impressive zero-shot performance on unseen tasks. So as a point of comparison, we evaluated the performance of several LLMs for zero-shot prompting, including Llama-2-70B (Touvron et al., 2023),

[5]We chose the number of 7,000 to make it comparable to the data generated by the pattern-based method.

Method	Macro F1	EE			Neutral		
		Pre	Rec	F1	Pre	Rec	F1
Llama-2	43.7	23.1	**95.3**	37.1	**97.2**	33.9	50.2
Falcon	65.8	36.8	79.1	50.2	94.3	71.6	81.4
GPT3.5	**70.2**	**44.0**	68.3	**53.5**	92.5	**81.9**	**86.9**
EEC$_{gold}$	83.5	73.2	72.1	72.6	94.2	94.5	94.4

Table 4: Zero-shot prompting and gold training results.

Method	Macro F1	EE			Neutral		
		Pre	Rec	F1	Pre	Rec	F1
GPT3.5	70.2	44.0	68.3	53.5	92.5	81.9	86.9
EEC with							
E_{PAT}	71.5	**68.0**	40.6	50.8	88.6	**96.0**	92.2
E_{LM}	74.7	52.4	69.2	59.6	93.1	86.8	89.9
$E_{LM} \times 2$	74.5	53.3	66.6	59.2	92.7	87.7	90.1
$E_{PAT} \cup E_{LM}$	**79.3**	62.1	**71.1**	**66.3**	**93.8**	91.0	**92.4**
EEC$_{gold}$	83.5	73.2	72.1	72.6	94.2	94.5	94.4

Table 5: Results with weakly labeled data only.

Falcon-180B (Penedo et al., 2023) and GPT3.5. Figure 3 shows the prompt template that we used.[6]

The first three rows of Table 4 shows the zero-shot prompting performance. The best model is GPT3.5, which achieves a macro F1 score of 70.2%. The highest F1 score for Embodied Emotion, however, is only 53.5%. This indicates that all models struggle to recognize embodied emotions.

The last row of Table 4 (EEC$_{gold}$) shows the performance of EEC trained with the gold training data (see Section 6.3). The supervised learning model achieves an F1 score of 83.5%, substantially outperforming the zero-shot prompting results.

6.2 Weak Supervision Results

Next, we train EEC using **only** weakly labeled data. We explored different sets of weakly labeled Embodied Emotion instances. Specifically, we trained EEC using:

E_{PAT} : The Embodied Emotion instances (7,162) produced by the pattern-based method.

E_{LM} : The Embodied Emotion instances (7,000) produced by the LM-based prompting method.

For all experiments, we use the Neutral instances generated by the LM-based prompting method, denoted by N_{LM}. In each experiment, we randomly selected instances from N_{LM} to enforce a distribution of 20% Embodied Emotion and 80% Neutral (to match the gold distribution). For each set of weakly labeled data, we then randomly selected 2,000 instances for validation and used the rest for training. Details of the model hyperparameters are provided in Appendix D.1.

Table 5 presents the results averaged across three runs. The first row shows the performance of zero-shot prompting with GPT3.5 once again, for the sake of comparison. Rows 2 to 6 show the performance of models trained with different sets of weakly labeled data. All of these models outperform zero-shot prompting. The E_{PAT} model achieves a macro F1 score of 71.5% , while the E_{LM} model achieves 74.7% F1 score. For the Embodied Emotion class, the E_{PAT} model has higher precision but the E_{LM} model has higher recall. This suggests that the E_{PAT} data is more precise while the E_{LM} data is more diverse.

Next, we tried adding more training data. The $E_{LM} \times 2$ row shows results when using twice as many Embodied Emotion instances (14,000) labeled by the prompting method, and twice as many Neutral instances. This model produces a macro F1 score of 74.5%, which is comparable to the E_{LM} model. This suggests that the value of this weakly labeled data source has maxed out.

Our next experiment trains EEC using both types of weakly labeled data together ($E_{PAT} \cup E_{LM}$). This training set contains 14,162 Embodied Emotion instances, with a corresponding balance of Neutral instances. Table 5 shows that training with both sets of data together produces a substantially better classifier, resulting in a F1 score of 79.3%. Importantly, note that training with 14k instances produced by two different methods yields much better results than training with 14k instances produced by the prompting method alone. These results suggest that the instances produced by the two methods are complementary.

The bottom row of Table 5 again shows the result of the model trained with gold data, for easy comparison. The model trained with only weakly labeled data ($E_{PAT} \cup E_{LM}$) performs nearly as well as the model trained with gold supervision (just 4.2 points lower in F1 score). We conclude that an embodied emotion classifier can be effectively trained using only weakly labeled data.

[6]We also experimented with few-shot prompting but it produced worse performance, which is reported in Appendix C.

Method	Macro F1	EE			Neutral		
		Pre	Rec	F1	Pre	Rec	F1
EEC$_{gold}$	83.5	**73.2**	72.1	72.6	94.2	**94.5**	94.4
$+weak$	**85.4**	72.9	**79.5**	**76.1**	**95.7**	93.9	**94.7**

Table 6: Using gold and weakly labeled data together.

6.3 The Best of Both Worlds: Exploiting Both Gold and Weakly Labeled Data

We also investigated whether the weakly labeled data could provide additional benefits when combined with gold labeled data. So we fine-tuned EEC using both the gold training data and the weakly labeled data together. Specifically, we used the best performing weakly labeled data: negative examples from N_{LM} and positive examples from $E_{PAT} \cup E_{LM}$. We used EEC fine-tuned with only gold data for comparison.

During training, we optimize the model with respect to the weighted cross entropy loss: $L = L_{gold} + \lambda L_{weak}$, where L_{gold} is the loss over the gold data, L_{weak} is the loss over the weakly labeled data and λ is a hyperparameter. We provide the model hyperparameters in Appendix D.2.

Table 6 shows the model performance averaged across three runs. The model trained with only gold data (row 1) yields a macro F1 of 83.5%. When the weakly labeled data is added (row 2), the model improves to achieve an F1 score of 85.4%. This improvement is mainly due to a large increase of 7.4 points in recall of Embodied Emotion (72.1% $\rightarrow$ 79.5%). Overall, the addition of the weakly labeled data helps the model recognize many more instances of embodied emotion with nearly the same precision.

7 Analysis

We present several analyses to better understand the behavior of our embodied emotion classifier.

Ablation Study In Section 6.3, we showed that combining the gold training data with weakly labeled data improves the performance of our classifier. So we further investigated how the different sources of weakly labeled data (E_{LM} and E_{PAT}) impact the model. Table 7 shows the performance when we remove one source at a time. Removing either source decreases performance, particularly on the recall for embodied emotions which drops from 79.5% down to 74.9% without E_{PAT} or to 76.3% without E_{LM}. These results reinforce the earlier observation that the weakly labeled data produced by the two different methods seem to be complementary and so using them together is beneficial.

Method	Macro F1	Embodied Emotion			Neutral		
		Pre	Rec	F1	Pre	Rec	F1
All	**85.4**	**72.9**	**79.5**	**76.1**	**95.7**	**93.9**	**94.7**
- E_{PAT}	83.9	72.3	74.9	73.5	94.8	**93.9**	94.3
- E_{LM}	84.1	71.7	76.3	73.9	95.0	93.7	94.3

Table 7: The effects of removing E_{PAT} or E_{LM} from the weakly labeled data, one at a time.

Body Part Frequency Analysis Some body parts are mentioned much more frequently than others (see Appendix A for frequency counts). We expect the classifier to generalize across body parts to some degree, but some body parts are fundamentally different than others (e.g., eyebrows vs. spine) so we also expect substantially different language around different body parts. We did an analysis to see how the amount of training data for a specific body part correlates with performance on instances of that body part.

Figure 4: F1 scores based on body part frequency.

We partitioned the 55 body parts[7] into two groups: 27 high frequency body parts with ≥ 20 training examples and 28 low frequency body parts with < 20 training examples. Figure 4 plots the F1 score for each body part on the y-axis. Overall, there is a strong correlation between training frequency and performance: most high frequency body parts show high F1 scores, with a few exceptions. The low frequency body parts typically have only one or a few instances in the test set so their performance is volatile, but most perform poorly. This analysis strongly suggests that producing additional training data for low frequency body parts would likely further improve our model.

Error Analysis We manually analyzed the errors of the best EEC in Table 6 and categorized them

[7]The dataset has 56 body parts but one did not occur in the test set.

False Negative Embodied Emotion
(a) He glared up at Ianto. "Thought I told you I didn't want to see your face." Ianto bit his **lip**.
(b) The tragedies that I had brought to my family and friends broke into fragments and stabbed me, as though they were taking revenge ... My **chest** began to tighten ...
False Positive Embodied Emotion
(c) The doctor there told me, "you are having a heart attack even as we speak." My **heart** arrested twice, I was shocked four times.
(d) Eames choked and gasped for air, his **head** already pounding from where he hit the other man.

Table 8: Error cases.

into two types. The first error type is **false negative Embodied Emotion**. For most cases of this error, we suspect the classifier failed because it cannot recognize the causal relationship between an emotional experience in the preceding context and the physical reaction. We show two examples in the upper portion of Table 8. For instance, "*lip*" in (a) is involved in embodied emotion as the biting results from the negative conversation in the preceding context. The second error type is **false positive Embodied Emotion**. Most cases of this error mention body parts involved in physical disorders. The bottom portion of Table 8 shows two examples. Polysemy may explain why the classifier is confused by many of these cases. For example, in (c) the word "shocked" refers to an electrical shock (presumably defibrillation), but it also commonly refers to an reaction to a surprise. Similarly, in (d) the word "gasped" simply refers to sharp breathing in this case, but it commonly refers to an emotional response.

8 Conclusion

Our work presents the first study on recognizing expressions of embodied emotion in natural language. We created a dataset that contains 7,300 body part mentions with human annotation for this task, which can be found at https://github.com/yyzhuang1991/Embodied-Emotions. We performed extensive experiments to show that this task is challenging for existing emotion recognition methods.

We also presented two methods to automatically produce a large set of weakly labeled instances, one pattern-based method that extracts manner expressions with explicit emotion words, and one prompting method that exploits a large language model. We showed that the weakly labeled data can be used to train an effective embodied emotion classifier, and that combining it with gold data yields a better classifier than using gold data alone.

9 Limitations

To create our dataset, we randomly selected sentences containing body part words from a large corpus, which led to an unbalanced distribution of body parts. For example, the top 10 body parts account for over 64% (4,720/7,300) of all sentences in our dataset. However, we did not manually force an even distribution of body parts as we wanted to keep the naturally occurring distribution, which led to the long tail issue. Another limitation associated with data diversity comes from the choice of the text corpus. We used text from a web blog corpus, which we believe is suitable for identifying daily life events and emotions. But the dataset inherently presents a bias toward personal narratives, which might not adequately capture the embodied emotions in other domains. For example, our body part list does not contain medical terminology and human anatomy names. Finally, in this project, we only identify if an expression is embodied emotion, without explicitly identifying the type of emotion. For future work, we want to conduct a more nuanced study by concentrating on the exact emotions associated with the embodied emotion expression, as compared to the existing emotion classification task.

References

Hassan Alhuzali and Sophia Ananiadou. 2021. SpanEmo: Casting multi-label emotion classification as span-prediction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2021)*.

Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 2005. Emotions from text: Machine learning for text-based emotion prediction. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (EMNLP 2005)*.

Magda B. Arnold. 1968. *The nature of emotion*. Penguin.

Lisa Barrett, Kristen Lindquist, Gerard Semin, and Eric Smith. 2008. The embodiment of emotion. In *Embodied Grounding: Social, Cognitive, Affective, and Neuroscientific Approaches*, pages 237–262. Cambridge University Press.

Lawrence W. Barsalou and Katya Wiemer-Hastings. 2005. Situating abstract concepts. In *Grounding Cognition: The Role of Perception and Action in Memory, Language, and Thinking*, pages 129–163. Cambridge University Press.

Felix Casel, Amelie Heindl, and Roman Klinger. 2021. Emotion recognition under consideration of the emotion component process model. In *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)*, pages 49–61, Düsseldorf, Germany. KONVENS 2021 Organizers.

Gustave Cortal, Alain Finkel, Patrick Paroubek, and Lina Ye. 2023. Emotion recognition based on psychological components in guided narratives for emotion regulation. In *Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 72–81, Dubrovnik, Croatia. Association for Computational Linguistics.

Charles Darwin. 1872. *The Expression of Emotions in Man and Animals*. John Murray.

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 (ACL 2020)*.

Lingjia Deng and Janyce Wiebe. 2014. Sentiment Propagation via Implicature Constraints. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2014)*.

Lingjia Deng and Janyce Wiebe. 2015. Joint Prediction for Entity/Event-Level Sentiment Analysis using Probabilistic Soft Logic Models. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2019)*.

Haibo Ding and Ellen Riloff. 2018. Weakly Supervised Induction of Affective Events by Optimizing Semantic Consistency. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI 2018)*.

Paul Ekman. 1972. Universals and cultural differences in facial expressions of emotion. In *Nebraska symposium on motivation*. University of Nebraska Press.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & Emotion*, 6:169–200.

Nico H. Frijda. 1986. *The Emotions*. Cambridge University Press.

Aniruddha Ghosh, Guofu Li, Tony Veale, Paolo Rosso, Ekaterina Shutova, John Barnden, and Antonio Reyes. 2015. SemEval-2015 task 11: Sentiment analysis of figurative language in Twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*.

Roberto González-Ibáñez, Smaranda Muresan, and Nina Wacholder. 2011. Identifying sarcasm in Twitter: A closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (NAACL 2011)*.

Chenyang Huang, Amine Trabelsi, Xuebin Qin, Nawshad Farruque, Lili Mou, and Osmar Zaïane. 2021. Seq2Emo: A sequence to multi-label emotion classification model. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2021)*.

William James. 1884. What is an emotion? *Mind*, 9(34):188–205.

Burton Kevin, Java Akshay, and Soboroff Ian. 2009. The ICWSM 2009 Spinn3r Dataset. In *Third Annual Conference on Weblogs and Social Media (ICWSM 2009)*, San Jose, CA. AAAI.

Burton Kevin, Java Akshay, and Soboroff Ian. 2011. The ICWSM 2011 Spinn3r Dataset. In *Proceedings of the Annual Conference on Weblogs and Social Media (ICWSM 2011)*. AAAI.

Evgeny Kim and Roman Klinger. 2019. An analysis of emotion communication channels in fan-fiction: Towards emotional storytelling. In *Proceedings of the Second Workshop on Storytelling*, pages 56–64, Florence, Italy. Association for Computational Linguistics.

George Lakoff and Mark Johnson. 1999. *Philosophy in the flesh: the embodied mind and its challenge to Western thought*. Basic Books.

Zhengyan Li, Yicheng Zou, Chong Zhang, Qi Zhang, and Zhongyu Wei. 2021. Learning implicit sentiment in aspect-based sentiment analysis with supervised contrastive pre-training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021)*.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)*.

George A. Miller. 1992. Wordnet: A lexical database for english. In *Human Language Technology - The Baltic Perspectiv*.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 task 1: Affect in tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval 2018)*.

Saif Mohammad, Ekaterina Shutova, and Peter Turney. 2016. Metaphor as a medium for emotion: An empirical study. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics (*SEM 2016)*.

Saif Mohammad and Peter Turney. 2010. Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Myriam Munezero, Tuomo Kakkonen, and Calkin Montero. 2011. Towards automatic detection of antisocial behavior from texts. In *Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2011)*.

Paula M. Niedenthal. 2007. Embodying emotion. *Science*, 316:1002 – 1005.

W. Gerrod Parrott. 2001. Emotions in social psychology: essential readings. Psychology Press.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

James Pennebaker, Martha Francis, and Roger Booth. 1999. In *Linguistic inquiry and word count: LIWC 2001*. Mahway: Lawrence Erlbaum Associates.

Jesse Prinz. 2004. Embodied emotions. In Robert C. Solomon, editor, *Thinking About Feeling: Contemporary Philosophers on Emotions*, pages 44–58. Oxford University Press.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 Task 4: Sentiment Analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval 2017)*.

Klaus R. Scherer. 2005. What are emotions? and how can they be measured? *Social Science Information*, 44(4):695–729.

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 2004)*.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*.

Yuan Zhuang, Tianyu Jiang, and Ellen Riloff. 2020. Affective event classification with discourse-enhanced self-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*.

Yuan Zhuang and Ellen Riloff. 2020. Exploring the role of context to distinguish rhetorical and information-seeking questions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 306–312, Online. Association for Computational Linguistics.

Yuan Zhuang and Ellen Riloff. 2023. Eliciting affective events from language models by multiple view co-prompting. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3189–3201, Toronto, Canada. Association for Computational Linguistics.

A Appendix: Breakdown Statistics of CHEER

Table 9 shows the frequencies of different body parts in CHEER.

B Appendix: Reproducing Emotion Classifiers

SpanEmo: We used the released code at https://github.com/hasanhuz/SpanEmo to train a SpanEmo model over the SemEval2018 dataset.

Seq2Emo: We used the released code at https://github.com/chenyangh/Seq2Emo to train Seq2Emo over the GoEmotions dataset. Although Seq2Emo was also reported to achieve the SOTA performance over the SemEval2018 dataset, we report the performance of Seq2Emo that was trained over GoEmotions, as it performs better over our dataset.

GoEmotion-BERT: As there is no released code, we developed code to train a BERT-base model over the GoEmotion dataset and reported its performance over our dataset.

Aff-BERT: We used the released code and pre-trained weights at https://github.com/yyzhuang1991/DEST. Since the model was designed to take event tuples, we only experiment with the *Event* strategy.

In our reproduction, all reproduced models achieved performance that is comparable to the reported performance in the corresponding paper.

B.1 Experimental Results

We present the performance of emotion classifiers with all four input strategies in Table 10. For SpanEmo, Seq2Emo and GE-BERT, the *Window* strategy consistently has a higher macro F1 score than *Multi-sent* and *Sent*. From the breakdown of Embodied Emotion, we see that the *Window* strategy has higher precision but lower recall than the other two strategies. This is probably because contexts of a smaller scope contain less irrelevant emotion information such as the emotions of other people. The *Window* strategy also outperforms the *Event* strategy except for SpanEmo, mainly because the *Event* strategy has lower recall of Embodied Emotion. This is probably because the emotion classifiers could not recognize emotion for event phrases. Indeed, Aff-BERT achieves a much higher recall of Embodied Emotion than other emotion classifiers with the *Event* strategy, since it is trained to recognized affective polarity for event phrases. However, its recall of Neutral is much lower. This is not surprising, because events that are affective are not necessarily Embodied Emotion, such as events that mention physical disorders (e.g., "*My leg hurts from the exercise*").

Overall, the best method is GE-BERT (*Window*) among all models, but it only achieves 58.2% macro F1 score. This implies that our task is not a trivial subset of standard emotion classification.

C Appendix: Prompting Large Language Model with Few Shots

Much work has shown that few-shot learning outperforms zero-shot learning. So we also explored few-shot learning by prompting the LLM with the best zero-shot performance, GPT3.5, with 4, 8 and 16 demonstration examples randomly drawn from the gold training set. In each experiment, 50% of the examples were Embodied Emotion. We performed 3 random runs for each experiment and show the average performance in the lower portion of Table 11. Surprisingly, all few-shot learning results are worse than the zero-shot learning result, and using more demonstration examples leads to worse performance.

D Appendix: Hyperparameters

For all experiments in the unsupervised and supervised learning settings, we set the maximum sequence length in BERT to be 256 and the batch size to 16. We also used the AdamW optimizer with a linear schedule and a warmup rate of 0.1. Before gradient descent, we clipped the gradient norm using the threshold of 1.0.

D.1 Appendix: Learning without Gold Data

We observed in our early experiments that varying the number of training epochs and the learning rate did not have a significant impact. So we trained the model for 10 epochs with a learning rate of 1e-5 for all experiments.

D.2 Appendix: Learning with Gold Data and Weakly Labeled Data

For the number of training epochs, we tried 5 and 10. For the learning rate, we searched through the set of (1e-5, 2e-5, 3e-5). For the weight hyperparameter λ, we searched through the range from 0.1

Table 9: Frequencies of different body parts in CHEER.

head (953), eye (853), hand (691), face (559), heart (384), foot (306), arm (267), leg (255), mouth (251), back (201), shoulder (170), finger (168), ear (143), stomach (138), knee (132), lip (129), chest (129), neck (126), throat (115), nose (111), tooth (104), brain (102), cheek (95), skin (92), tongue (60), ankle (56), lung (55), hip (48), toe (44), thumb (40), forehead (39), spine (31), belly (30), nail (29), jaw (29), eyebrow (28), chin (28), palm (28), wrist (27), waist (25), nerve (25), elbow (22), fist (21), thigh (20), muscle (18), heel (18), rib (15), temple (13), eyelid (13), bone (12), skull (11), vein (11), calf (10), knuckle (8), abdomen (7), forearm (5)

Method	**Macro F1**	**EE**			**Neutral**		
		Pre	Rec	F1	Pre	Rec	F1
SpanEmo							
Multi-sent	26.7	18.1	**92.1**	30.3	**89.0**	13.2	23.0
Sent	32.0	18.1	83.9	29.8	86.3	21.2	34.1
Window	37.6	18.4	74.4	29.5	85.4	31.1	45.6
Event	45.2	18.4	53.7	27.4	83.9	50.4	63.0
Seq2Emo							
Multi-sent	52.3	21.0	33.3	25.7	84.2	74.0	78.8
Sent	53.7	23.3	22.8	23.1	84.0	91.4	87.5
Window	54.4	28.8	16.7	21.2	84.1	91.4	87.6
Event	51.0	24.7	9.4	13.7	83.3	**94.0**	**88.3**
GE-BERT							
Multi-sent	52.6	21.6	36.2	27.1	84.5	72.6	78.1
Sent	54.0	23.2	30.1	26.2	84.5	79.3	81.8
Window	**58.2**	**31.0**	30.3	30.6	85.6	85.9	85.7
Event	53.5	28.4	14.4	19.1	83.8	92.5	87.9
Aff-BERT$_{Event}$	50.3	21.7	56.1	**31.3**	86.4	57.8	69.3

Table 10: Evaluating emotion classification models.

Method	**Macro F1**	**EE**			**Neutral**		
		Pre	Rec	F1	Pre	Rec	F1
4-shot	**69.8**	**48.3**	57.1	**51.0**	90.9	**86.9**	**88.7**
8-shot	68.9	43.9	62.3	50.9	**91.5**	83.1	87.0
16-shot	58.5	29.7	**65.6**	40.3	90.6	67	76.6

Table 11: Results of prompting GPT3.5 with few shots.

to 1.0 with an increment of 0.1. We then selected the hyperparameters that performed the best over the gold validation set.