# Disambiguating Emotional Connotations of Words Using Contextualized Word Representations

**Akram Sadat Hosseini**
University of Stuttgart
Stuttgart, Germany
akram-sadat.hosseini@ki.uni-stuttgart.de

**Steffen Staab**
University of Stuttgart
Stuttgart, Germany
University of Southampton
Southampton, UK
Steffen.Staab@ki.uni-stuttgart.de

## Abstract

Understanding emotional nuances in written content is crucial for effective communication; however, the context-dependent nature of language poses challenges in precisely discerning emotions in text. This study contributes to the understanding of how the emotional connotations of a word are influenced by the sentence context in which it appears. Leveraging the contextual understanding embedded in contextualized word representations, we conduct an empirical investigation to (i) evaluate the varying abilities of these representations in distinguishing the diverse emotional connotations evoked by the same word across different contexts, (ii) explore potential biases in these representations toward specific emotions of a word, and (iii) assess the capability of these representations in estimating the number of emotional connotations evoked by a word in diverse contexts. Our experiments, utilizing four popular models—BERT, RoBERTa, XL-Net, and GPT-2—and drawing on the GoEmotions and SemEval 2018 datasets, demonstrate that these models effectively discern emotional connotations of words. RoBERTa, in particular, shows superior performance and greater resilience against biases. Our further analysis reveals that disambiguating the emotional connotations of words significantly enhances emotion identification at the sentence level.

## 1 Introduction

Understanding the emotional nuances conveyed by words is crucial for effective communication. This insight enhances the design of conversational agents that emulate human empathy, enabling responses that accurately reflect the emotions conveyed by word choice (Raji and de Melo, 2021). Psycholinguistics leverages this understanding to identify depression and suicidality, where specific word usage on social media posts may indicate underlying distress (Aragón et al., 2019). Moreover, comprehending these emotional subtleties alongside literal meanings of words can deepen second language comprehension for non-native speakers (Dewaele, 2010).

In cognitive linguistics, the concept of `emotional connotation` refers to the emotion attributed to a specific word, transcending its explicit meaning (Stubbs, 1995). Take, for instance, the word 'damn', which is rated by humans as *anger* (Mohammad and Kiritchenko, 2018), likely stemming from its frequent use in expressions of anger. However, a word may take on various emotional connotations depending on the context in which it appears. Consider the word 'damn' in the following sentences sourced from the GoEmotions dataset (Demszky et al., 2020):

`S1.` Wash your damn hands. [Anger]
`S2.` Damn [NAME] is KILLING it. [Joy]
`S3.` I damn near broke down! [Sadness]
`S4.` Damn, that's dark here! [Fear]

In `S1`, the word 'damn' expresses anger, while in `S2`, it is used in a positive context to convey joy. Both `S3` and `S4` exemplify its usage in other negative contexts.

Research on determining the emotional connotations of lexical items has typically utilized crowdsourcing methods, leading to the development of diverse lexicons of words with predefined emotions (Hofmann et al., 2020). These lexicons, however, provide static and generalized ratings for words, regardless of the context in which they are used (De Bruyne et al., 2022). Additionally, despite attempts to ensure consistency in word ratings through anchoring, implicit biases may persist in the rating process (Semeraro et al., 2023). Efforts to address these limitations have mainly focused on distinguishing the polarity of words (Hellrich et al., 2019). In a domain-specific corpus (soccer), Braun et al. (2022) relied on human judgments to measure the differences between the positivity and negativ-

ity of words with and without a sentence context. Moreover, adopting automatic methods as an alternative to manual acquisition has been limited to extending the lexicon's word coverage (Sedoc et al., 2017) or developing domain-specific polarity lexicons (Hamilton et al., 2016).

Recently, contextualized word representations, exemplified by BERT (Devlin et al., 2019), have been frequently evaluated on word relatedness benchmarks, such as word sense disambiguation (Wiedemann et al., 2019), which is the task of identifying the correct sense of a word's usage from a fixed inventory of sense identifiers (Hadiwinoto et al., 2019). Studies on textual emotion analysis have utilized these representations particularly for *sentence-level* emotion classification tasks (Chen et al., 2023; Fan et al., 2022; Huang et al., 2021; Alhuzali and Ananiadou, 2023).

What is particularly intriguing about contextualized word representations is their ability to generate unique embeddings for a word based on its context (Saravia et al., 2018). Our objective is to leverage this property of contextualized word representations to disentangle the emotional connotations of words that evoke various emotions within different sentence contexts. Let $W$ be the set of all words, where $w \in W$ represents a target word, which evokes different emotions depending on its surrounding context. $\mathcal{S}$ is the set of all possible sentences, and $\mathbb{N}$ is the set of natural numbers representing the position in a sentence where we aim to analyze the emotional connotation of the word $w$. The function $f$, $f(S, i) = e$, categorizes the dominant emotional connotation of $w$ at position $i$ in sentence $S$. The signature of this function is given by $f : \mathcal{S} \times \mathbb{N} \rightarrow \mathcal{E}$, where $\mathcal{E}$ is the set of all possible emotion categories.

We conduct an empirical investigation to:

(i) Evaluate the varying abilities of contextual word representations in distinguishing the diverse emotional connotations evoked by the same word across different contexts;

(ii) Explore the existence of potential biases in these representations toward specific emotional connotations of a word; in this context, bias refers to the likelihood of models incorrectly associating a word linked to $emotion_k$ with $emotion_j$;

(iii) Assess the capability of these representations in estimating the number of emotions a word can evoke in various contexts; and

(iv) Investigate the impact of disambiguating the emotional connotations of words on the accuracy of sentence-level emotion classification.

Focusing on emotional words that evoke various emotions across diverse contexts, we obtain contextualized representations of these words within emotion-annotated sentences in the GoEmotions and the SemEval 2018 (Mohammad et al., 2018) datasets. We then cluster these representations and assess the alignment degree between the resulting clusters and the emotions of the words in question. Our analysis of various models—BERT, RoBERTa, XLNet, and GPT-2—showcases their capability to capture the emotional connotations of words. We find that not all models are equally effective in discerning these emotional nuances. Our findings also reveal biases towards specific emotions in these representations, with different models exhibiting biases towards different emotions for a given word. Moreover, our experiments indicate that disambiguating the emotional connotations of a small number of words significantly improves the accuracy of sentence-level emotion classification.

## 2 Related work

Textual emotion recognition has typically involved either the utilization of lexicons—lists of words with pre-assigned emotions—without the need for extensive labeled data (Semeraro et al., 2023), or contextualized word representations, known for their domain-agnostic adaptability, when sufficient labeled data is available (Öhman et al., 2020).

Methods that rely on lexicons view texts as word collections and use word ratings from lexicons for emotion identification (Ma et al., 2018; Hosseini and Staab, 2023). However, the static nature of these word ratings limits a comprehensive understanding of emotions, as they disregard contextual nuances (De Bruyne et al., 2022). For instance, in a domain-specific corpus (soccer), Braun et al. (2022) demonstrated that pragmatic and semantic shifts in context can significantly influence word polarity in lexicons. To address this limitation, researchers often explore the identification of negations, diminishers, and intensifiers (Reitan et al., 2015; Hutto and Gilbert, 2014), or they develop domain-specific lexicons (Amir et al., 2015), which have mainly focused on distinguishing polarity of words in a specific domain (Hellrich et al., 2019).

Recent methods in textual emotion analysis have increasingly leveraged contextual word representations like BERT, particularly for sentence-level

emotion classification (Alhuzali and Ananiadou, 2023; Li et al., 2021; Mao et al., 2023). These methods enhance model training by fine-tuning these embeddings with emotion-labeled datasets. For instance, Batbaatar et al. (2019) applied these representations in a Convolutional Neural Network to discern semantic relationships between words, and Kassner and Schütze (2020) focused on refining the understanding of contradictory sentiment words within these representations for binary sentiment classification. Some studies have integrated emotional lexicons into these representations, enabling these models to achieve a more nuanced understanding of emotional words (Aduragba et al., 2021; Ke et al., 2020; Wang et al., 2020). For example, Sosea and Caragea (2021) proposed a pretraining objective for BERT, which increases masking probabilities for emotional words in sentences using emotion lexicons, while Zhou et al. (2020) developed a BERT model from scratch using Yelp and Amazon reviews by increasing the masking probability for positive and negative words. However, these approaches exhibit high sensitivity to both the training corpus and the lexical resources employed (Shah et al., 2023), which suffer from ambiguity (Wang et al., 2021). Certain studies, such as Wang and Zong (2021), focused on semantic role labeling for emotions, modeling the semantics and interrelatedness of emotion labels by learning representations for emotion classes from annotated data. However, these representations do not generalize to other datasets and label formats (Campagnano et al., 2022).

Previous studies have primarily utilized contextualized representations for *sentence-level* emotion classification tasks. In contextualized word representations, each input word is represented as a vector dependent on the context of its occurrence (Saravia et al., 2018). This approach captures both semantic and syntactic nuances within the surrounding context of words, rendering these models particularly intriguing for investigating the emotional connotations of words across diverse contexts. This paper exploits these representations to conduct an empirical study, aiming to scrutinize their efficacy in distinguishing different emotional connotations evoked by the same word in various contexts. To achieve this, we adopt a clustering-based approach, wherein the representation vectors of the word, obtained from emotion-annotated sentences, are clustered using a Gaussian Mixture Model. Further, we evaluate potential biases in different representation models toward certain emotional connotations of words and assess whether clustering is a viable method for predicting the number of emotions a word can evoke in diverse contexts.

## 3 Methodology

To investigate the effectiveness of contextualized word representations in discerning the various emotional connotations of words that evoke different emotions depending on the sentence context, we propose a method comprising the following steps:
1. *Target Word Identification:* Identify a subset of emotional words within an emotion lexicon that evoke diverse emotions across different contexts.
2. *Sample Sentence Extraction:* Retrieve sentences featuring the target words from emotion-annotated resources to compile a representative set of instances showcasing the words in diverse contexts.
3. *Contextualized Representation Generation:* Obtain contextualized representation vectors for the target words in the set of sample sentences.
4. *Word Representation Clustering.* Apply clustering to the contextualized representations using a Gaussian Mixture Model (GMM) and find a mapping between the resulting clusters and the emotions of target words that maximizes the overall number of accurate matches, with the match rate serving as the evaluation metric.

The next sections detail the target word identification phase, the word representations used in our study, and the clustering of these representations.

### 3.1 Target Word Identification

The task of identifying emotional words that can evoke multiple emotions in different contexts relies on two foundational assumptions:

**Assumption 1:** The subset of emotional words eliciting diverse emotions is significantly smaller than the set of words maintaining consistent emotional connotations across various contexts (Wang et al., 2021; Gollapalli et al., 2020).

**Assumption 2:** The emotional connotation of a word can be inferred by analyzing its frequency within a corpus of annotated text. If an emotional word frequently appears in sentences expressing a specific emotion, it is reasonable to deduce that it is commonly employed to convey that emotion (Liu, 2022; Hosseini, 2017).

To identify words that evoke various emotions based on context, the inherent emotionality of a word is a prerequisite for our study. We utilized the

NRC-Affect lexicon (Mohammad and Kiritchenko, 2018), a well-established resource in emotion analysis, to extract emotional words from annotated datasets. This lexicon, annotated manually, comprises 4,192 English words and their associations with four basic emotions (*anger, fear, sadness, and joy*) with scores ranging from 0 to 1. It encompasses common English terms and terms prevalent on social media platforms.

The initial step involved extracting words from the NRC-Affect lexicon that were present in various emotional classes of the annotated datasets. We then calculated the proportion of the word's frequency in each emotional class to its total frequency across all sentences, as shown in (1):

$$\text{Proportion}(w, e) = \frac{\text{freq}(w, e)}{\sum_e \text{freq}(w, e)} \quad (1)$$

Here, $\text{freq}(w, e)$ denotes the frequency of candidate word $w$ in emotion category $e$, where $e$ represents each emotion category in the dataset. Formula (1) generates values between 0 and 1, with the sum equal to 1, indicating the normalized frequency of extracted words in distinct categories, irrespective of dataset size. For a word $w$ to be considered a target word, a minimum normalized frequency of 0.2 in each emotion category is required. This criterion reduces noise in the identification process and strikes a balance between being stringent enough to filter out less relevant words while remaining practical for analysis.

We refined the target word selection process further by requiring a minimum occurrence in the 25 annotated sentences for each emotion. For example, the word 'crazy' met this criterion, appearing in 75 sentences expressing *anger* and 40 sentences expressing *joy*. In contrast, 'abortion' did not meet the criteria, as it appeared in sentences expressing various emotions (anger, fear, and sadness) but lacked the required number of annotated sentences per emotion. Setting a minimum occurrence criterion ensures the identified words have a robust presence in the dataset. To ensure a balanced distribution of sentences across emotions and prevent bias towards more frequent emotion classes, we imposed a maximum limit of 100 sentences per emotion. In line with Assumption 2, we associated the emotions of the identified target words with the emotion expressed within sentences.

### 3.2 Contextual Representation Generation

This section provides an overview of the contextualized word representations used in this paper, e.g., BERT, RoBERTa, XLNet, and GPT-2. These models were selected based on their prevalent use in sentiment analysis and text emotion analysis (Chen et al., 2023; Fan et al., 2022; Mao et al., 2023). They embody a broad spectrum of transformer architectures, with unique objectives and pre-training methods. The coverage includes bidirectional models (BERT, RoBERTa, XLNet) and a unidirectional model (GPT-2), incorporating various language modeling approaches such as masked language modeling and autoregressive language modeling. Table 1 summarizes their differences in corpus size, parameters, embedding dimensions, and layers. We used publicly available pre-trained versions of these models specified by 'bert- large-uncased,' 'roberta-large,' 'xlnet-base-cased' and 'gpt2' on Hugging Face.

- **BERT** (Devlin et al., 2019) employs masked language modeling and next-sentence prediction to generate bidirectional text representations, considering both preceding and succeeding context.
- **RoBERTa** (Liu et al., 2019), built on BERT's architecture, omits the next-sentence prediction task and introduces dynamic masking, which generates unique masking patterns for each sentence during training rather than using a fixed masked token.
- **XLNet** (Yang et al., 2019) is an autoregressive language model that employs permutation-based training to predict random tokens in both directions, allowing for bidirectional context capture.
- **GPT-2** (Radford et al., 2019) is a unidirectional autoregressive language model that employs the Transformer decoder architecture for its generative pre-training, specializing in predicting the next word in a sentence by considering preceding words.

| Model | Params. | Corpus Size | Tokenization | Dims. | Layers |
|---|---|---|---|---|---|
| BERT | 340M | 16GB | WordPiece | 1024 | 24 |
| RoBERTa | 355M | 160GB | Byte-Pair | | |
| XLNet | 340M | 158GB | SentencePiece | | |
| GPT-2 | 345M | 40GB | Byte-Pair | | |

Table 1: Details of contextualized word representations used in this study.

### 3.3 Word Representation Clustering

We utilized the Gaussian Mixture Model (GMM) from scikit-learn for clustering the generated contextualized word vectors, selecting the 'spherical'

covariance type, which assumes equal diagonal elements in a diagonal covariance matrix. The GMM can adapt to clusters with diverse shapes and sizes while employing a probabilistic method for clustering (Melnykov and Maitra, 2010). Through an optimization strategy, we then identified a mapping between the resulting clusters and the emotions of the target words that maximize the overall number of accurate matches. Using the match rate as the evaluation metric leads to a more refined measure of clustering quality.

The match rate quantifies the alignment between the resulting clusters and the emotions of target words. We determined this rate by constructing a contingency table with `pandas.crosstab` (denoted as $C$), where each cell $C_{ij}$ counts the instances in cluster $i$ associated with emotion label $j$. Initially, we assigned an emotion label to each cluster based on the predominant emotion of the instances within that cluster, following a majority voting principle. Subsequently, we refined this alignment by employing the Hungarian algorithm (Kuhn, 1955) to establish an optimal one-to-one mapping between clusters and target words' emotions. This optimization seeks a permutation $\pi$ that minimizes mismatch costs, thereby maximizing the alignment between clusters and emotions. The match rate was then calculated by normalizing the sum of correctly matched labels, according to the optimal mapping $\pi$, by the total count of instances $n$, as follows:

$$\text{Match Rate} = \frac{\sum_{i=1}^{k} C_{i,\pi(i)}}{n} \quad (2)$$

Here, $k$ denotes the number of clusters, and $\pi(i)$ represents the label matched with cluster $i$ through the optimal matching.

## 4 Experiments

In this section, we first investigate the ability of various contextualized word representations to distinguish between the different emotional connotations evoked by the same word in different contexts (Section 4.2). Then, we explore the presence of emotional biases in these representations (Section 4.3). Finally, we evaluate the accuracy of these representations in quantifying the range of emotions elicited by each word (Section 4.4).

### 4.1 Datasets

We used the GoEmotions and SemEval 2018 datasets, sourced from Reddit and Twitter, respectively, as emotion-annotated datasets.

• **GoEmotions** (Demszky et al., 2020) is the largest manually annotated dataset of 58k English Reddit comments from popular subreddits. At least three raters assessed each comment, resulting in significant inter-rater agreement. Comments range from 3 to 30 tokens, with a median length of 12 tokens. We utilized the version of the dataset annotated for six emotions: joy, anger, fear, sadness, disgust, and surprise.

• **SemEval 2018** (Mohammad et al., 2018) comprises 10,983 tweets annotated for 11 emotions: anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise, and trust. At least seven raters assessed each tweet, ensuring reliable annotation with strong inter-rater correlation. The tweets range from 1 to 36 tokens, with a median length of 16 tokens.

We followed the procedure outlined in Section 3.1 for target word identification. In the GoEmotions dataset, we identified 133 words with an average of 2.5 distinct emotions, and in the SemEval dataset, 113 words with an average of 2.3 emotions per word. For evaluation, we selected 90 and 80 words from the GoEmotions and SemEval datasets, respectively, as the test set, reserving the remaining words for parameter fine-tuning in the development set. The emotional labels for these words were assigned based on the emotions expressed in the sentences, including anger, fear, sadness, and joy. Emotions like *surprise*, although present in the datasets, did not meet the criteria outlined in Section 3.1 and were thus excluded from our analysis. We then retrieved example sentences associated with these words from the datasets, with an average of 58.37 annotated sentences per emotion in GoEmotions and 32.17 in SemEval.

### 4.2 Emotional Connotations Distinction

This section investigates the effectiveness of various contextualized representations in identifying varied emotional connotations evoked by a single word in different contexts. To ensure the robustness of our experiments, we conducted five clustering trials with different random seeds and selected the result with the highest likelihood.

Figure 1 presents a comparative analysis of macro-average match rates across all words for individual layers within four representation models on the development set, using the GoEmotions and SemEval datasets. This empirical evidence reveals

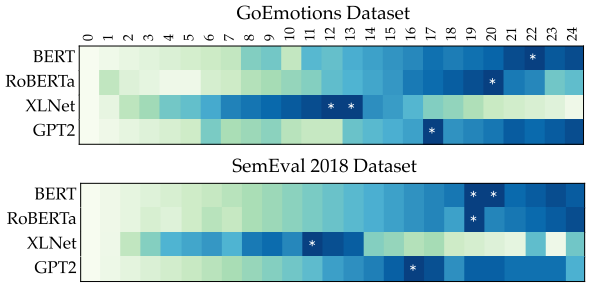


Figure 1: Layer-wise comparison of macro-average match rates across models on the development set for GoEmotions and SemEval datasets. Blue shading highlights the best-performing layers, marked with ⋆.

| Model | GoEmotions | | SemEval 2018 | |
|---|---|---|---|---|
| | Dev | Test | Dev | Test |
| BERT | 0.694 (22) | 0.689 (22) | 0.631 (20) | 0.623 (20) |
| RoBERTa | **0.718** (20) | **0.707** (20) | **0.665** (19) | **0.656** (19) |
| XLNet | 0.662 (13) | 0.639 (13) | 0.643 (11) | 0.631 (11) |
| GPT-2 | 0.618 (17) | 0.598 (17) | 0.595 (16) | 0.587 (16) |

Table 2: Macro-average match rates of the highest-scoring development layers (in brackets) and their corresponding test set scores across models. Bold values represent the top scores for each datasets.

significant variation in layer effectiveness across different models, indicating that selecting an optimal layer is model-specific. The final layer (Layer 24), typically associated with encoding semantic knowledge, consistently underperforms across all models. A hierarchical performance pattern is observed in BERT and RoBERTa models, with higher match rates in the upper layers. In contrast, the XLNet and GPT-2 models perform best in layers closer to the middle rather than in the final layer.

Table 2 presents the macro-average match rates of the top-performing layers during development and their corresponding scores on the test set. Given the diverse contextualization approaches, objectives, and pre-training strategies of the models in question, significant variations were noted in their ability to discern the emotional nuances conveyed by the same words across different contexts. RoBERTa emerged as the leading model in terms of scores on both datasets, underscoring its superior ability to differentiate emotional connotations of words and position them into distinct embedding space regions. Following RoBERTa, XLNet and BERT—both employing bidirectional architectures—demonstrated strong performance. Conversely, GPT-2, which operates on a unidirectional autoregressive language model framework, recorded the lowest scores on both datasets.

### 4.3 Bias Analysis

This section investigates the presence of biases toward specific emotional connotations of a word in contextualized representations, aiming to enhance our understanding of their behavior in distinguishing different emotions of the same word.

We aim to measure $Bias(w, j)$, related to the $j$-th emotion ($emotion_j$), for a word $w$ that evokes multiple emotions ($n$). This involves assessing $c_{ij}$, the count of instances where the correct label is $emotion_i$ but is erroneously identified as $emotion_j$ ($i \neq j$). First, we determine the extent of bias from emotion $i$ to emotion $j$ ($bias_{ij}$) by normalizing $c_{ij}$, dividing it by the total number of instances gold-labeled as $emotion_i$, denoted as $\sum_j c_{ij}$. We then compute the overall bias towards a specific emotion, $Bias(w, j)$, as follows:

$$Bias(w, j) = \frac{1}{n-1} \sum_{i=1, i \neq j}^{n} \left( \frac{c_{ij}}{\sum_j c_{ij}} \right) \quad (3)$$

The value of $Bias(w, j)$ represents the likelihood of models incorrectly identifying a word associated with $emotion_k$ as $emotion_j$ when $k \neq j$ (Loureiro et al., 2021). This value ranges from 0 to 1, where a value close to 1 indicates a stronger bias towards $emotion_j$. We calculate the maximum bias value ($\max(Bias(w, j))$ towards different emotions of a word, with $j$ ranging across the emotions associated with the word ($j \in [1, n]$). Table 3 shows the average of these maximum bias values across all words for the four models. Consistent with the findings in Section 4.2, our analysis indicates RoBERTa is more robust against biases, maintaining a bias value below 0.3 in both datasets.

Table 4 presents the average $Bias(w, j)$ scores from equation 3 for different emotions across all words. This breakdown analysis reveals biases in word representations toward specific emotions, with variations observed across different models. Although models generally exhibit similar behavior, they do not uniformly exhibit identical bias toward the same emotions. For instance, in the GoEmotions dataset, RoBERTa is biased toward *Anger*, whereas XLNet and GPT-2 lean towards *Joy*. Moreover, the models consistently show low biases, below 0.2, towards *Fear* and *Sadness* emotions.

The radar charts in Figure 2 illustrate biases towards different emotions in several representative cases. For instance, the words 'Freak', 'Damn', and

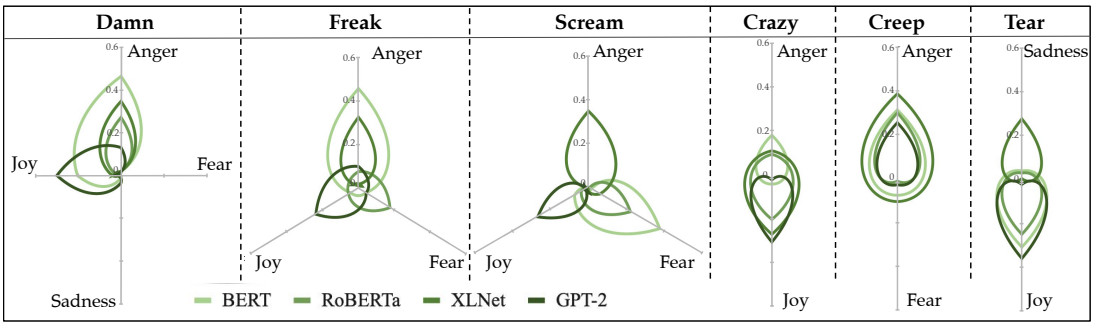


Figure 2: Analysis of bias towards different emotions for a few representative cases.

| Dataset | BERT | RoBERTa | XLNet | GPT-2 |
|---|---|---|---|---|
| GoEmotions | 0.310 | **0.257** | 0.331 | 0.396 |
| SemEval 2018 | 0.380 | **0.286** | 0.349 | 0.422 |

Table 3: Average bias values across models for the GoEmotions and SemEval datasets.

| Model | Anger | Fear | Sadness | Joy |
|---|---|---|---|---|
| *GoEmotions Dataset* | | | | |
| BERT | **0.283** | 0.091 | 0.186 | 0.251 |
| RoBERTa | **0.267** | 0.116 | 0.103 | 0.233 |
| XLNet | 0.241 | 0.118 | 0.191 | **0.281** |
| GPT-2 | 0.287 | 0.141 | 0.149 | **0.322** |
| *SemEval 2018 Dataset* | | | | |
| BERT | **0.387** | 0.135 | 0.118 | 0.285 |
| RoBERTa | **0.297** | 0.054 | 0.152 | 0.276 |
| XLNet | **0.353** | 0.114 | 0.112 | 0.275 |
| GPT-2 | 0.390 | 0.145 | 0.148 | **0.415** |

Table 4: Breakdown analysis of bias values toward various emotions across different models and datasets.

'Creep' exhibit a bias towards *Anger*, and 'Scream' is skewed toward *Fear*, showcasing a preference for the more prevalent emotions of these words. This reflects biases present during the models' pre-training, meaning they encountered the target word with the prevalent emotion more frequently than with other emotions. Thus, they may overlook the word's emotional nuance in varied contexts. For example, these models associated the word 'damn' with more negative emotions due to its frequent co-occurrence with words like 'fuck' and 'shit' during pre-training, potentially missing the word's positive connotations in different contexts. Moreover, different models exhibit varying biases for the same word; for instance, RoBERTa shows a bias toward *Anger* for 'Freak', whereas GPT-2 leans towards *Joy*. Similarly, for the word 'Crazy', BERT, RoBERTa, and GPT-2 tend to misclassify *Anger* as *Joy*, whereas XLNet does the opposite.

### 4.4 Number of Emotions Estimation

In prior experiments, we analyzed words that evoke diverse emotions and provided the Gaussian Mixture Model with the number of emotions present in sample sentences. The current experiment aims to assess the models' accuracy in estimating the number of emotional connotations of words. By including words that elicit only a single emotion alongside those evoking multiple emotions, we enhance the robustness of our evaluation.

For the implementation of this experiment, we matched the number of additional words to the quantity used in Section 4.2. This approach resulted in parameter tuning with 86 and 66 words in the development set for GoEmotions and SemEval datasets, respectively. For the test set evaluation, we utilized 180 words for the GoEmotions and 160 words for the SemEval dataset.

We utilized an adjusted version of the Bayesian Information Criterion (ABIC) (Schwarz, 1978) as the criterion for model selection to determine the optimal number of clusters, which align with the number of emotions elicited by each word. The ABIC fine-tunes the model for the best fit to the data by considering both model complexity and mitigating overfitting, as specified by the formula:

$$\text{ABIC} = c \cdot p \cdot \ln(N) - 2\ln(\hat{L}) \qquad (4)$$

Here, $\hat{L}$ denotes the maximum likelihood of the model, $N$ is the sample size, $p$ is the number of model parameters, and $c$ is a constant to adjust the penalty term, $c \cdot p \cdot \ln(N)$. The penalty term penalizes model complexity based on the number of parameters and discourages excessive increases in the number of clusters. We increment $c$ from 1 in 0.1 steps until the total number of emotions and the estimated number of clusters are as close as possible

| Model | GoEmotions | | | SemEval 2018 | | |
|---|---|---|---|---|---|---|
| | $\rho$ | Accuracy | RMSE | $\rho$ | Accuracy | RMSE |
| BERT | 0.513 | 0.572 | 1.133 | 0.231 | 0.503 | 1.254 |
| RoBERTa | **0.648** | **0.617** | **1.002** | **0.437** | **0.569** | **1.142** |
| XLNet | 0.359 | 0.521 | 1.211 | 0.327 | 0.519 | 1.239 |
| GPT-2 | 0.189 | 0.465 | 1.291 | 0.151 | 0.434 | 1.303 |

Table 5: Comparison of different models in estimating the number of emotions using Spearman's ($\rho$), accuracy, and Root Mean Square Error (RMSE).

in the development set. For the GoEmotions and SemEval datasets, the optimal $c$ values were identified as 3.6 and 3.2, respectively. In the GMM, each cluster encompasses a mean, a spherical covariance matrix, and a mixture weight. The parameter count for the GMM is given by $p = [N_c \times (D + 2)] - 1$, where $N_c$ is the number of clusters, and $D$ the data dimensionality. The term $(D + 2)$ accounts for the mean and covariance parameters, and subtracting 1 corrects for the constraints, ensuring the sum of parameters equals 1 for mixture weights (Murphy, 2012; Yamada et al., 2021).

Table 5 presents the performance of various models on estimating the number of emotions and clusters in both datasets, using accuracy, Spearman's rank correlation coefficient ($\rho$), and root mean square error (RMSE) metrics. RMSE quantifies the error magnitude between estimated cluster counts and the actual emotion counts per word. The findings indicate that RoBERTa surpasses other models in accurately estimating the number of emotions for over 60% of the words analyzed. RoBERTa achieved the lowest RMSE and the highest $\rho$ values—0.648 and 0.437 for the GoEmotions and SemEval datasets, respectively—suggesting a strong alignment between the actual number of emotions and model estimates. Figure 3 in the Appendix A further illustrates RoBERTa's performance through confusion matrices, analyzing its emotion count estimates for words with a single emotion and those with context-dependent multiple emotions.

## 5 Sentence-level Emotion Classification

This section explores the impact of disambiguating the emotional connotation of words that evoke different emotions depending on the context, on the accuracy of sentence-level emotion detection. We evaluate sentences containing at least one identified target word, as those without these words remain unaffected. Sentences are divided into stratified training (80%) and test (20%) splits based on emotions through random sampling.

Our initial experiments involve comparing the original NRC-Affect lexicon and its modified versions in a *before-and-after* manner. Here, *modified lexicon* refers to the disambiguation of emotional connotations associated with target words in the original NRC-Affect lexicon, achieved by utilizing various contextualized word representations. The probability values from the Gaussian Mixture Model indicate the extent to which each instance of a target word belongs to each of the GMM clusters, which have been mapped to specific emotions. We computed the average probability for all instances of a target word within an emotion's cluster to ascertain its disambiguated ratings. For example, while the original lexicon associated 'damn' exclusively with *anger*, with a score of 0.7, the modified lexicon provides a nuanced view of the different emotional connotations—joy, sadness, and fear, in addition to anger—that 'damn' evokes across various contexts in the GoEmotions dataset. Building on the lexicon-based classifier design outlined in (De Bruyne et al., 2022), we utilized the information from both the original NRC-Affect lexicon and its modified versions as features in a logistic regression classifier for emotion prediction, detailed in Appendix B. Table 6 presents the results using F1-macro scores, demonstrating substantial improvements with the modified lexicons compared to the original. This underscores the crucial role of addressing ambiguous emotional words and considering context in determining their emotional connotations for accurate emotion classification.

| Method | GoEmotions | SemEval 2018 |
|---|---|---|
| Original NRC-Affect | 0.324 | 0.361 |
| Modified NRC-Affect using | | |
| ├ BERT | 0.377 | 0.396 |
| ├ RoBERTa | **0.382** | **0.408** |
| ├ XLNet | 0.372 | 0.406 |
| └ GPT-2 | 0.356 | 0.390 |

Table 6: The F1-macro scores for sentence-level emotion classification using lexicons.

| Method | GoEmotions | SemEval 2018 |
|---|---|---|
| BERT | 0.593 | 0.532 |
| RoBERTa | *0.621* | *0.561* |
| XLNet | 0.614 | 0.543 |
| GPT-2 | 0.461 | 0.503 |
| RoBERTa + Original NRC-Affect | 0.631 | 0.569 |
| RoBERTa + Modified NRC-Affect (RoBERTa) | **0.636** | **0.573** |

Table 7: The F1-macro scores for sentence-level emotion classification using pre-trained models.

In the second series of experiments, we evaluated the ability of pre-trained transformer models—BERT, RoBERTa, XLNet, and GPT-2—to classify emotions in sentences with target words. We applied a uniform set of hyperparameters across all models, adhering to the settings recommended by Demszky et al. (2020): four epochs, a batch size of 16, and a learning rate 5e-5. As expected, the results in Table 7 demonstrate that these models significantly outperformed the lexicon-based approach, which depended solely on lexicon features, with RoBERTa achieving the highest F1-macro scores across both datasets. Building on prior research that indicates incorporating lexicon information into linguistic models further enhances the understanding of emotional nuances in pre-trained models (Baziotis et al., 2018), we integrated features derived from the original and top-performing modified lexicons as auxiliary inputs into the highest-performing pre-trained model. Specifically, we concatenated the auxiliary features with the output vector from the last hidden layer of the pre-trained model, appending them to the sequence embedding since the features aggregate across the entire text. The concatenated vector was then fed into the final decision-making layer, and we adjusted the dimensionality of the final layer to accommodate the additional inputs. Our findings, detailed in the second set of entries in Table 7, revealed that including modified lexicon features, in addition to the models, enhances classification performance beyond what is achieved with original lexicon information. Appendix C further discusses the enhanced ability of RoBERTa, compared to other models in discerning emotions.

**Overall performance.** Table 8 compares our method, using the modified NRC-Affect lexicon and RoBERTa embeddings, with state-of-the-art approaches across the entire GoEmotions and SemEval datasets, covering sentences both with and without identified target words. We compare our results with various models, such as the TCS model, which uses dual BiLSTM networks for tweet encoding (Meisheri and Dey, 2018); the DATN model, which employs a dual attention mechanism within a transfer learning setup (Yu et al., 2018); the BERT+DK, that integrates domain knowledge into BERT (Ying et al., 2019); the Seq2Emo, which leverages a bi-directional decoder in a sequence-to-emotion framework without relying on external data(Huang et al., 2021); and the UCCA-GAT and Dep-GAT models (Ameer et al., 2023) that integrate semantic and syntactic information into graph attention networks via Universal Conceptual Cognitive Annotation and dependency trees, respectively. Our approach surpasses most competing models, though it falls slightly behind the UCCA-GAT on the SemEval dataset. These findings highlight the efficacy of contextualized representations to disambiguate emotional connotations of words and adapt to varying contexts, thereby enhancing emotion detection at the sentence level.

| Method | F1-macro |
|---|---|
| *SemEval 2018 Dataset* | |
| TCS Research (Meisheri and Dey, 2018) | 0.530 |
| DATN (Yu et al., 2018) | 0.544 |
| BERT-Large + DK (Ying et al., 2019) | 0.563 |
| Seq2Emo (Huang et al., 2021) | 0.519 |
| UCCA-GAT (Ameer et al., 2023) | **0.600 (1)** |
| Dep-GAT (Ameer et al., 2023) | **0.578 (3)** |
| RoBERTa + Modified NRC-Affect (RoBERTa) | **0.583 (2)** |
| *GoEmotions Dataset* | |
| BERT (Demszky et al., 2020) | **0.640 (2)** |
| UCCA-GAT (Ameer et al., 2023) | **0.639 (3)** |
| Dep-GAT (Ameer et al., 2023) | 0.611 |
| RoBERTa + Modified NRC-Affect (RoBERTa) | **0.653 (1)** |

Table 8: Comparison of our method using modified NRC-Affect lexicon and RoBERTa embeddings with state-of-the-art approaches. Rankings (1), (2), and (3) denote the top three results.

## 6 Conclusion

In this study, we have explored disentangling the emotional connotations of words within diverse sentence contexts, leveraging contextualized word representations. We evaluated these representations' ability to differentiate the diverse emotions of words, identify potential biases in predicting emotional connotations, and accurately estimate the multiplicity of words' emotional connotations. Our methodology involved clustering based on contextualized representations of words that evoke different emotions in various contexts and assessing the alignment between the generated clusters and the words' emotions. Our evaluation of BERT, RoBERTa, XLNet, and GPT-2 models revealed that contextualized representations can effectively disambiguate the emotional connotations of words, with RoBERTa showing superior performance and greater resilience against biases. Further analysis indicated that addressing a small subset of ambiguous emotional words and considering the context in determining their emotional connotations are crucial for accurately determining sentence emotion.

## 7 Limitations

The empirical results presented in this paper highlighted that many commonly used linguistic models can significantly improve word emotion induction methods. However, our experiments were conducted exclusively on English-language datasets. Consequently, the effectiveness of the proposed method in diverse corpora and multilingual resources remains to be determined. Additionally, we employed the NRC-Affect lexicon as a resource to identify target emotional words that evoke different emotions depending on the context. However, this lexicon may not encompass all emotional words, such as emerging slang terms in social media. The inclusion of a more comprehensive spectrum of emotional words should be a priority in future research. These investigations will be essential for evaluating the applicability of our method across different languages and are expected to advance us toward the goal of automatically constructing high-quality emotional lexical resources with broader linguistic coverage for under-resourced languages or specific domains.

## References


Olanrewaju Tahir Aduragba, Jialin Yu, Alexandra I Cristea, and Lei Shi. 2021. Detecting fine-grained emotions on social media during major disease outbreaks: health and well-being before and during the covid-19 pandemic. In *AMIA annual symposium proceedings*, volume 2021, page 187. American Medical Informatics Association.

Hassan Alhuzali and Sophia Ananiadou. 2023. Improving textual emotion recognition based on intra- and inter-class variations. *IEEE Transactions on Affective Computing*, 14(2):1297–1307.

Iqra Ameer, Necva Bölücü, Grigori Sidorov, and Burcu Can. 2023. Emotion classification in texts over graph neural networks: Semantic representation is better than syntactic. *IEEE Access*, 11:56921–56934.

Silvio Amir, Ramon F. Astudillo, Wang Ling, Bruno Martins, Mario J. Silva, and Isabel Trancoso. 2015. INESC-ID: A regression model for large scale Twitter sentiment lexicon induction. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 613–618, Denver, Colorado. Association for Computational Linguistics.

Mario Ezra Aragón, Adrián Pastor López Monroy, Luis Carlos González-Gurrola, and Manuel Montes. 2019. Detecting depression in social media using fine-grained emotions. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 1481–1486.

Erdenebileg Batbaatar, Meijing Li, and Keun Ho Ryu. 2019. Semantic-emotion neural network for emotion recognition from text. *IEEE Access*, 7:111866–111878.

Christos Baziotis, Athanasiou Nikolaos, Alexandra Chronopoulou, Athanasia Kolovou, Georgios Paraskevopoulos, Nikolaos Ellinas, Shrikanth Narayanan, and Alexandros Potamianos. 2018. NTUA-SLP at SemEval-2018 task 1: Predicting affective content in tweets with deep attentive RNNs and transfer learning. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 245–255, New Orleans, Louisiana. Association for Computational Linguistics.

Lisa Beinborn and Yuval Pinter. 2023. Analyzing cognitive plausibility of subword tokenization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4478–4486, Singapore. Association for Computational Linguistics.

Nadine Braun, Martijn Goudbeek, and Emiel Krahmer. 2022. Affective words and the company they keep: Studying the accuracy of affective word lists in determining sentence and word valence in a domain-specific corpus. *IEEE Transactions on Affective Computing*, 13(3):1440–1451.

Cesare Campagnano, Simone Conia, and Roberto Navigli. 2022. SRL4E – Semantic Role Labeling for Emotions: A unified evaluation framework. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4586–4601, Dublin, Ireland. Association for Computational Linguistics.

Chih Yao Chen, Tun Min Hung, Yi-Li Hsu, and Lun-Wei Ku. 2023. Label-aware hyperbolic embeddings for fine-grained emotion classification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10947–10958, Toronto, Canada. Association for Computational Linguistics.

Luna De Bruyne, Pepa Atanasova, and Isabelle Augenstein. 2022. Joint emotion label space modeling for affect lexica. *Computer Speech Language*, 71:101257.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep

bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Jean-Marc Dewaele. 2010. *Emotions in multiple languages*. Springer.

Shuai Fan, Chen Lin, Haonan Li, Zhenghao Lin, Jinsong Su, Hang Zhang, Yeyun Gong, JIan Guo, and Nan Duan. 2022. Sentiment-aware word and sentence level pre-training for sentiment analysis. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4984–4994, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sujatha Das Gollapalli, Polina Rozenshtein, and See-Kiong Ng. 2020. ESTeR: Combining word co-occurrences and word associations for unsupervised emotion detection. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1043–1056, Online. Association for Computational Linguistics.

Christian Hadiwinoto, Hwee Tou Ng, and Wee Chung Gan. 2019. Improved word sense disambiguation using pre-trained contextualized word representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5297–5306, Hong Kong, China. Association for Computational Linguistics.

William L. Hamilton, Kevin Clark, Jure Leskovec, and Dan Jurafsky. 2016. Inducing domain-specific sentiment lexicons from unlabeled corpora. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 595–605, Austin, Texas. Association for Computational Linguistics.

Johannes Hellrich, Sven Buechel, and Udo Hahn. 2019. Modeling word emotion in historical language: Quantity beats supposed stability in seed word selection. In *Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 1–11, Minneapolis, USA. Association for Computational Linguistics.

Jan Hofmann, Enrica Troiano, Kai Sassenberg, and Roman Klinger. 2020. Appraisal theories for emotion classification in text. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 125–138, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Akram Sadat Hosseini. 2017. Sentence-level emotion mining based on combination of adaptive meta-level features and sentence syntactic features. *Engineering Applications of Artificial Intelligence*, 65:361–374.

Akram Sadat Hosseini and Steffen Staab. 2023. Emotional framing in the spreading of false and true claims. In *Proceedings of the 15th ACM Web Science Conference 2023*, WebSci '23, page 96–106, New York, NY, USA. Association for Computing Machinery.

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*, pages 4717–4724, Online. Association for Computational Linguistics.

C. Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1):216–225.

Nora Kassner and Hinrich Schütze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7811–7818, Online. Association for Computational Linguistics.

Pei Ke, Haozhe Ji, Siyang Liu, Xiaoyan Zhu, and Minlie Huang. 2020. SentiLARE: Sentiment-aware language representation learning with linguistic knowledge. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6975–6988, Online. Association for Computational Linguistics.

Harold W Kuhn. 1955. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97.

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*, pages 246–256, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Bing Liu. 2022. *Sentiment analysis and opinion mining*. Springer Nature.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Daniel Loureiro, Kiamehr Rezaee, Mohammad Taher Pilehvar, and José Camacho-Collados. 2021. Analysis and evaluation of language models for word sense disambiguation. *Computational Linguistics*, 47:387–443.

Yukun Ma, Haiyun Peng, and Erik Cambria. 2018. Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive lstm. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Rui Mao, Qian Liu, Kai He, Wei Li, and Erik Cambria. 2023. The biases of pre-trained language models: An empirical study on prompt-based sentiment analysis and emotion detection. *IEEE Transactions on Affective Computing*, 14(3):1743–1753.

Hardik Meisheri and Lipika Dey. 2018. Tcs research at semeval-2018 task 1: Learning robust representations using multi-attention architecture. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 291–299.

Volodymyr Melnykov and Ranjan Maitra. 2010. Finite mixture models and model-based clustering.

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*, pages 1–17, New Orleans, Louisiana. Association for Computational Linguistics.

Saif Mohammad and Svetlana Kiritchenko. 2018. Understanding emotions: A dataset of tweets to study interactions between affect categories. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Kevin P Murphy. 2012. *Machine learning: a probabilistic perspective*. MIT press.

Emily Öhman, Marc Pàmies, Kaisla Kajava, and Jörg Tiedemann. 2020. XED: A multilingual dataset for sentiment analysis and emotion detection. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6542–6552, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Shahab Raji and Gerard de Melo. 2021. Guilt by association: Emotion intensities in lexical representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9911–9917, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Johan Reitan, Jørgen Faret, Björn Gambäck, and Lars Bungum. 2015. Negation scope detection for Twitter sentiment analysis. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 99–108, Lisboa, Portugal. Association for Computational Linguistics.

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. CARER: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697, Brussels, Belgium. Association for Computational Linguistics.

Gideon Schwarz. 1978. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

João Sedoc, Daniel Preoţiuc-Pietro, and Lyle Ungar. 2017. Predicting emotional word ratings using distributional representations and signed clustering. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 564–571, Valencia, Spain.

Alfonso Semeraro, Salvatore Vilella, Saif Mohammad, Giancarlo Ruffo, and Massimo Stella. 2023. Emoatlas: An emotional profiling tool merging psychological lexicons, artificial intelligence and network science.

Sapan Shah, Sreedhar Reddy, and Pushpak Bhattacharyya. 2023. Retrofitting light-weight language models for emotions using supervised contrastive learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3640–3654, Singapore. Association for Computational Linguistics.

Tiberiu Sosea and Cornelia Caragea. 2021. emlm: A new pre-training objective for emotion related tasks. In *Annual Meeting of the Association for Computational Linguistics*, pages 286–293.

Michael Stubbs. 1995. Collocations and semantic profiles: On the cause of the trouble with quantitative studies. *Functions of Language*, 2(1):23–55.

Shuai Wang, Guangyi Lv, Sahisnu Mazumder, and Bing Liu. 2021. Detecting domain polarity-changes of words in a sentiment lexicon. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3657–3668, Online. Association for Computational Linguistics.

Shuo Wang, Aishan Maoliniyazi, Xinle Wu, and Xiaofeng Meng. 2020. Emo2vec: Learning emotional embeddings via multi-emotion category. *ACM Trans. Internet Technol.*, 20(2).

Xiangyu Wang and Chengqing Zong. 2021. Distributed representations of emotion categories in emotion space. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2364–2375.

Gregor Wiedemann, Steffen Remus, Avi Chawla, and Chris Biemann. 2019. Does bert make any sense? interpretable word sense disambiguation with contextualized embeddings. *arXiv preprint arXiv:1909.10430*.

Kosuke Yamada, Ryohei Sasano, and Koichi Takeda. 2021. Verb sense clustering using contextualized word representations for semantic frame induction.

In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4353–4362, Online. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Wenhao Ying, Rong Xiang, and Qin Lu. 2019. Improving multi-label emotion classification by integrating both general and domain-specific knowledge. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 316–321, Hong Kong, China. Association for Computational Linguistics.

Jianfei Yu, Luís Marujo, Jing Jiang, Pradeep Karuturi, and William Brendel. 2018. Improving multi-label emotion classification via sentiment classification with dual attention transfer network. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1097–1102, Brussels, Belgium. Association for Computational Linguistics.

Jie Zhou, Junfeng Tian, Rui Wang, Yuanbin Wu, Wenming Xiao, and Liang He. 2020. Sentix: A sentiment-aware pre-trained model for cross-domain sentiment analysis. In *Proceedings of the 28th international conference on computational linguistics*, pages 568–579.

## A Appendix

The confusion matrices presented in Figure 3 depict the analysis of how the RoBERTa model estimates the number of emotions for words with single emotions versus those with context-dependent multiple emotions. The model reliably identifies single-emotion words across datasets in most cases. However, it occasionally overestimates the number of emotions, suggesting 2 or 3 clusters. For words that elicit 2 or 3 emotions, the model provides reasonably accurate estimates but often underestimates the actual count, indicating a lower number of emotions. As the number of emotions increases to 4, the reliability of the model's estimations decreases, leading to a wider range of possibilities.

## B Appendix

We provide more details on the features utilized in the design of the lexicon-based classifier outlined in (De Bruyne et al., 2022). We employed statistical features of emotional words and a logistic regression classifier in the learning model for emotion prediction experiments. We trained

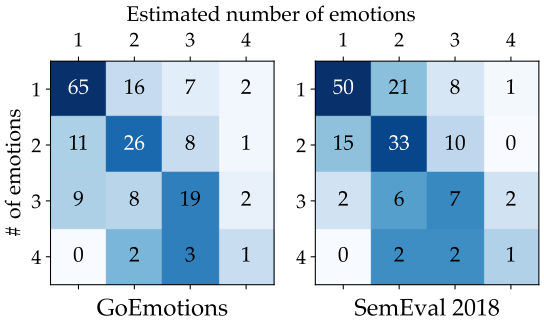


Figure 3: Confusion matrices for estimating the number of emotions using the RoBERTa model.

the classifier with the statistical features derived from both the original NRC-Affect lexicon and its modified versions. Given a sequence of words $s = (w_1, \ldots, w_k)$, statistical features quantify the proportion of a given emotion $e_i$ within the sentence $s$ as $P(s, e_i)$, calculated by:

$$P(s, e_i) = \frac{1}{k} \sum_{j=1}^{k} \phi_{e_i}(w_j) \quad (5)$$

where $\phi_{e_i}(w_j)$ represents the emotion score of the word $w_j$ for the emotion $e_i$, derived from the lexicons. Here, $e_i$ belongs to the set $\{e_{\text{anger}}, e_{\text{fear}}, e_{\text{sadness}}, e_{\text{joy}}\}$.

The logistic regression classifier uses a liblinear solver with L2 regularization and a regularization strength of $C = 1.0$. The choice of L2 regularization helps prevent model overfitting by penalizing the size of the coefficients, with $C = 1.0$ providing an optimal balance between regularization intensity and model complexity based on either empirical evidence. We deploy separate binary classifiers for each of the categories and aggregate the predictions afterward by selecting the highest probability, thereby identifying the most dominant emotion in the sentence.

## C Appendix

This appendix provides a detailed discussion on the superior performance of RoBERTa over other models—BERT, XLNet, and GPT-2—on the task of disambiguating emotional connotations. Notably, all models in our experiments were trained under identical conditions, using the same hyperparameters such as batch size, learning rate, and dataset sizes. This uniform setup ensures that any observed performance differences are due to architectural or training method variations. RoBERTa, the most

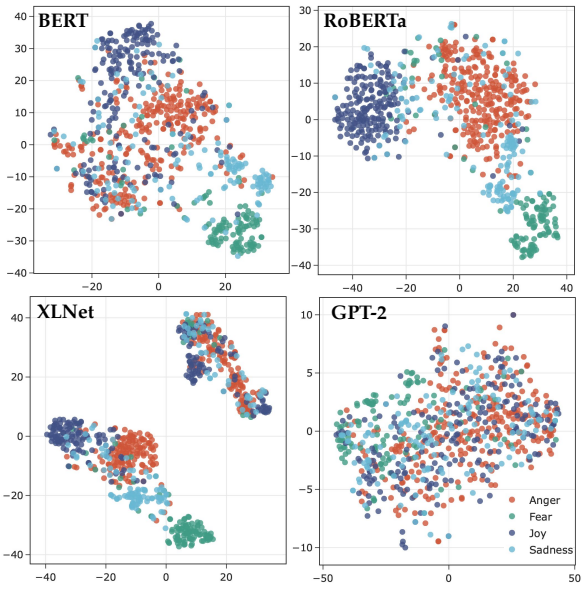


Figure 4: The t-SNE projection of BERT, RoBERTa, XLNet, and GPT-2 representations of the word *damn* in sentences expressing various emotions in GoEmotions dataset.

advanced transformer among the considered models, possesses the highest number of parameters (355M) as detailed in Table 1. Its dynamic masking technique, which alters mask patterns with each data pass, provides a training advantage over the static masking used by BERT. Additionally, the tokenization approach of models significantly impacts their performance. In our evaluation, tokenization was consistent with the method used during pre-training. RoBERTa employs Byte-Pair Encoding (BPE), which effectively captures frequent subword units compared to BERT's WordPiece or XLNet's SentencePiece. BPE constructs its vocabulary by merging frequently occurring character pairs or combinations, thus improving the capture of rare or out-of-vocabulary words (Beinborn and Pinter, 2023).

Additionally, empirical evidence from the layer-wise comparison of macro-average match rates in Section 4.2 revealed that the ability to capture emotional connotations varies significantly across layers of selected models, indicating that optimal layer selection is model-specific. RoBERTa consistently excelled in identifying the varying emotional connotations of words, particularly in its upper layers, which are typically associated with semantic knowledge encoding. Figure 4 showcases t-SNE projections of contextualized representations from the most effective layer of each model, using the word 'damn' in various sentences sourced from the GoEmotions dataset. These visualizations highlight the distinctive distribution of RoBERTa's representations, further emphasizing its ability to capture emotion evoked by 'damn' in each example.