SensoryT5: Infusing Sensorimotor Norms into T5 for Enhanced Fine-grained Emotion Classification

Yuhan Xia[1], Qingqing Zhao[2], Yunfei Long[1], Ge Xu[3], Jia Wang[4]

[1]School of Computer Science and Electronic Engineering, University of Essex
[2]Institute of Linguistics, Chinese Academy of Social Sciences
[3]College of Computer and Control Engineering, Minjiang University
[4]Department of Intelligent Science, Xi'an Jiaotong-Liverpool University
{yx23989, yl20051}@essex.ac.uk
zhaoqq@cass.org.cn, xuge@pku.edu.cn, Jia.Wang02@xjtlu.edu.cn

Abstract

In traditional research approaches, sensory perception and emotion classification have traditionally been considered separate domains. Yet, the significant influence of sensory experiences on emotional responses is undeniable. The natural language processing (NLP) community has often missed the opportunity to merge sensory knowledge with emotion classification. To address this gap, we propose SensoryT5, a neuro-cognitive approach that integrates sensory information into the T5 (Text-to-Text Transfer Transformer) model, designed specifically for fine-grained emotion classification. This methodology incorporates sensory cues into the T5's attention mechanism, enabling a harmonious balance between contextual understanding and sensory awareness. The resulting model amplifies the richness of emotional representations. In rigorous tests across various detailed emotion classification datasets, SensoryT5 showcases improved performance, surpassing both the foundational T5 model and current state-of-the-art works. Notably, SensoryT5's success signifies a pivotal change in the NLP domain, highlighting the potential influence of neuro-cognitive data in refining machine learning models' emotional sensitivity.

Keywords: emotion classification, sensory information, attention mechanism, pre-trained language model

1. Introduction

Affective computing stands at the intersection of technology and human emotions (Li et al., 2017), whereby sentiment analysis and emotion recognition are generally merged to give machines a semblance of human-like emotional understanding. Specifically, sentiment analysis (SA) seeks to decode the attitudes and viewpoints of opinion holders using computational methods (Lu et al., 2023), providing a coarse-grained categories of polarities: positive, negative, or neutral (Long et al., 2019b). Driven by recent advancements in deep learning and bolstered by vast labeled datasets, discriminating sentiments in standard contexts has become progressively more tractable. Cutting-edge models, including the likes of BERT (Devlin et al., 2018), XLNet (Yang et al., 2019), and the T5 (Raffel et al., 2020) series, have consistently set benchmarks, achieving high accuracies on an array of sentiment classification tasks.

By contrast, emotion analysis (EA) has received less notable results in recent years. One of the reasons is that different from SA offering a coarse-grained outlook, EA paints a detailed picture. That is, EA not only distinguishes between basic sentiments but also identifies nuanced emotions such as joy, anger, sadness, surprise, and among others (Ekman, 1992). Thus, the task of EA is complicated by the sheer variety of emotional categories. For in-

stance, distinguishing closely related emotions like "contentment" and "happiness" or "annoyance" and "anger" requires a discerning approach, especially when the medium is textual content. Thus, this study introduces a SensoryT5 model, tailored to infuse sensory data, which is cognitively more related to emotions and includes linguistically more enriched features, into neural architectures, to achieve a profound comprehension of emotions.

The relationship between emotion and perception/sensation has been verified repeatedly in various disciplines. From a neuroscientific perspective, emotion and sensory information are processed in an overlapping neural region, i.e., the amygdala (Šimić et al., 2021). Shifting the lens to psychology, emotion and perception are intertwined (Zadra and Clore, 2011). For example, the sense of taste shows an inherent link with reward and aversion mechanisms, such as sucrose being perceived as sweet and desirable, whereas quinine being recognized as bitter and repulsive (Yamamoto, 2008). In addition, emotion as a kind of interoception forms an indispensable part of human sensations, when a wide definition of sensory perception adopted (Connell et al., 2018; Lynott et al., 2020). In terms of the linguistic conceptualization of emotions, people more frequently use figurative language instead of literal emotion terms to convey emotions (Fainsilber and Ortony, 1987; Lee, 2018), and the conceptual metaphor EMOTION IS PERCEPTION is grounded

in abundant language usages to show that the human senses are fruitful sources for verbalizing emotions (e.g., sweet and bitter) (Lakoff and Johnson, 1980; Kövecses, 2019; Müller et al., 2021).

Given the intertwined relation between emotion and perception/sensation, this study posits that incorporating sensory information into a computational framework can capture the nuanced interplay between them, hence offering a reflection of intricate human affective understanding. Specifically, we utlize the Lancaster Sensorimotor Norms (Lynott et al., 2020), which include language-specific lexical properties representing the correlation between conceptualized lexical meanings and sensory modalities.

Our work boasts three pivotal advancements: (1) We introduce SensoryT5, an innovative architecture that enhances transformer-based fine-grained emotion classification models by seamlessly embedding sensory knowledge. Marking one of the pioneering endeavors, SensoryT5 is adapted at harmonizing both the nuances of contextual attention and the intricacies of sensory information-based attention. (2) The SensoryT5 leverages sensorimotor norms within transformer text classification frameworks, contributing to the ongoing efforts to incorporate neuro-cognitive data in NLP tasks. Thus, our work not only demonstrates the practical benefits of this integration in improving emotion classification tasks, but also encourages continued interdisciplinary dialogue and research between the domains of language processing and neuro-cognitive science. (3) Assessments across multiple real-world datasets pertinent to fine-grained emotion classification affirm that our approach amplifies the efficacy of pre-existing models considerably, even surpassing contemporary state-of-the-art methodologies on selected datasets. This endeavor underscores the value of cognition-anchored data in sculpting attention models. Our findings illuminate the untapped potential of sensory information in refining emotion classification, carving fresh prospects for exploration within the realm of affective computing in NLP.

2. Related work

2.1. Emotion analysis

Over recent years, the domain of pre-trained language models (PLMs) and large language models (LLMs) has witnessed marked advancements. Noteworthy developments include models like BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), GPT-3 (Brown et al., 2020), T5 (Raffel et al., 2020), PaLM (Chowdhery et al., 2022), LLaMA (Touvron et al., 2023) and ChatGPT (OpenAI, 2023). These models, through rigorous pre-training on vast text corpora using self-supervised learning, have the ability to autonomously generate intricate representations. This capability has significantly advanced the field, setting new benchmarks in numerous tasks, notably in sentiment analysis (Devlin et al., 2018; Zhang et al., 2023). An in-depth exploration by Zhang et al. (2023) elucidated the performances of LLMs in sentiment and emotion analysis tasks. The study has highlighted that, while LLMs excel over PLMs in few-shot learning scenarios, PLMs remain superior for more nuanced tasks that demand a deeper understanding of emotions or structured emotional data. Among the discussed models, T5 (Raffel et al., 2020) stands out due to its innovative 'text-to-text' transfer approach, in which every NLP challenge is remodelled as a text-to-text problem. Consequently, T5 frequently sets the state-of-the-arts in emotion analysis when utilized as the base model.

However, despite the considerable improvements made with these PLMs/LLMs, some research gaps remain relatively fulfilled. Present models, although they possess sophisticated neural architectures capable of discerning patterns from immense text datasets, often overlook the intricate nature of emotion—a dynamic interplay of cognitive and physiological responses triggered by various stimuli (Khare et al., 2023). Sensory perceptions, pivotal in shaping these responses, serve as the bedrock upon which our cognitive processes evaluate and generate emotions (Niedenthal and Wood, 2019). Integrating these models with sensory data can potentially elevate their performance, nudging them closer to approaching human-like comprehension. This presents a significant research opportunity: equipping already potent PLMs/LLMs with an element of sensory perception, an aspect they conventionally lack. With our proposed SensoryT5 model, our ambition is to fill this gap by synergizing the strengths of T5 and augmenting it with sensory knowledge, thereby enabling a deeper and more nuanced understanding of emotions.

2.2. Cognition-grounded resources: Sensorimotor norms

In recent years, there is an emergent trend that neuro-cognitive data and computational approaches are synergized in NLP studies. This interdisciplinary synergy unlocks new dimensions in understanding language, sentiment, and emotion, reflecting more accurately the human experience and mental processing. For instance, Long et al. (2019b) improved the attention model for sentiment analysis by incorporating a eye-tracking dataset. Chen et al. (2021) incorporated brain measurement data for modeling word embedding. Wan et al. (2023) demonstrated the superiority of neural

networks for metaphor detection by leveraging sensorimotor knowledge. These studies collectively underscore a broader shift in the field towards a more integrated approach to NLP. By weaving in neuro-cognitive data, researchers are equipping computational models with a richer and more intricate understanding of human language and cognition, which are often overlooked by traditional data-driven methods.

Given the intimate connection between emotion and perception as demonstrated in various studies reviewed in the last section, this study assumes that a cognitively and linguistically motivated representation of words in text based on sensorimotor knowledge would improve the performance of computational models for emotion analysis. That is not only because sensory inputs are crucial sources of emotions, but also because emotional responses are part of sensory perceptions for human beings.

Thus, this study utilizes Lynott et al. (2020)'s sensorimotor norms which encompass metrics of sensorimotor strengths (ranging from 0 to 5) of 39,707 concepts spanning six perceptual domains: touch, hearing, smell, taste, vision, and interoception, as well as five action effectors: mouth/throat, hand/arm, foot/leg, head (barring mouth/throat), and torso. To exploit this wealth of data, SensoryT5 is proposed to construct the sensorimotor vectors from these norms and to seamlessly embed them into the T5's decoder mechanism via an auxiliary attention layer. Positioned after the decoders, this sensory-centric attention layer is synergized with the decoder's output, producing an enriched representation brimming with sensory knowledge for words in text. Thus, SensoryT5 is adapted at simultaneously discerning contextual cues and sensory knowledge, allowing for a potent alignment of sensory nuances with contextual intelligence. This integration augments the model's efficacy in the fine-grained emotion classification.

3. Our proposed SensoryT5 model

In this section, we elaborate how our SensoryT5 model incorporates the sensory knowledge into the neural emotion classification framework. Specifically, sensory knowledge is infused into the T5 using an adapter approach built upon attention mechanisms. Moreover, the contextual and sensory information learning branches are amalgamated within a unified loss function to facilitate joint training. The overarching structure is depicted in Figure 1.

3.1. Preliminaries

Despite the relatively large size of the Lancaster Sensorimotor Norms, there are still many out-of-vocabulary words. Following the method proposed by Li et al. (2017), we use a word embedding model to regressively predict the sensory values of unknown words, aiming to obtain sensory values for out-of-vocabulary words.

Inputs and outputs The objective of emotion analysis is to determine and categorize opinions for a piece of texts following a defined label schema. Let D denote a collection of documents for emotion classification. Each document $d \in D$ is first tokenized into a word sequence with maximum length n, then the word embeddings w_i of these sequence are jointly employed to represent the document $d = w_1, w_2, ..., w_i, ..., w_n (i \in 1, 2, ..., n)$.

3.2. The core attention mechanism in T5

The word embeddings of these sequence $d = w_1, w_2, ..., w_i, ..., w_n (i \in 1, 2, ..., n)$ first enters the T5. Each layer of the encoder and decoder has a series of multi-head attention units. The multi-head attention mechanism for the final decoder layer can be represented using the following equation:

$$
\begin{aligned}
V_d &= \mathsf{MultiHead}(Q_0, K_0, V_0) \\
&= [\mathsf{head}_1, \mathsf{head}_2, ..., \mathsf{head}_i] W_O
\end{aligned}
\tag{1}
$$

Where each head is computed as:

$$
\begin{aligned}
\mathsf{head}_i &= \mathsf{Attention}(Q_0 W_i^Q, K_0 W_i^K, V_0 W_i^V) \\
&= \mathsf{softmax}\left(\frac{(Q_0 W_i^Q)(K_0 W_i^K)^T}{\sqrt{d_k}} \right) V_0 W_i^V
\end{aligned}
\tag{2}
$$

W_i^Q, W_i^K, and W_i^V are weight matrices that are learned during the training process. They are used to project the input queries (Q), keys (K), and values (V) to different sub-spaces. Q_0, K_0, and V_0 are derived from the output of the penultimate decoder layer. Additionally, following the common practice for text classification with T5, we employ a zero-padding vector as the sole input for the decoder. The result V_d is the output of the T5 decoder, imbued with context-aware attention. Both V_d and K_0 will be utilized in section 3.4 for integration with sensory knowledge.

3.3. Sensory information transformation for T5 integration

We project the Lancaster Sensorimotor Norms into a sensory word vector space. Each word is linked with a six-dimensional vector representing sensory scores across six perceptual modalities (auditory, gustatory, haptic, interoceptive, olfactory and visual dimensions). For a word w, its sensory vector is denoted as $s(w) = [s_1, s_2, ..., s_6]$.

To enable effective integration into the T5-large, we use two linear transformations followed by a
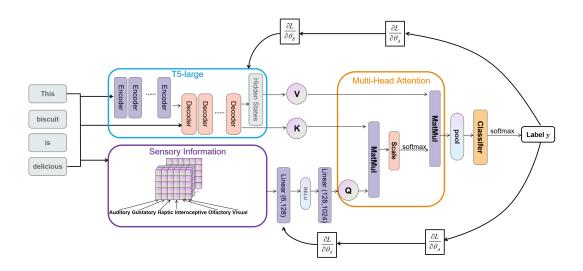
Figure 1: An overview of SensoryT5. Blue box shows a T5 process of deep learning, while purple box describing sensory information is quantified and passed into the T5.

ReLU activation function to map the sensory vectors to the same dimension as the T5-large's word embeddings. Given a T5-large model with an embedding dimension of 1024, the transformation process can be formally described as:

$$\mathbf{h}_1 = \mathsf{ReLU}(\mathbf{W}_1 s(w) + \mathbf{b}_1) \qquad (3)$$

$$s'(w) = \mathbf{W}_2 \mathbf{h}_1 + \mathbf{b}_2 \qquad (4)$$

where $\mathbf{W}_1 : \mathbf{R}^6 \to \mathbf{R}^{128}$ and $\mathbf{W}_2 : \mathbf{R}^{128} \to \mathbf{R}^{1024}$ are two linear transformation matrices and $\mathbf{b}_1$, $\mathbf{b}_2$ are the respective bias terms. The shapes of the two weight matrices W_1 and W_2 are respectively $(6, 128)$ and $(128, 1024)$. The output h_1 of the first linear layer is a vector of shape $(1, 128)$, and the output $s'(w)$ of the second linear layer is a vector of shape $(1, 1024)$. After the transformation, the sensory vector $s'(w)$ is projected into the same semantic space as the features generated by T5-large. The output vector $s'(w)$, with V_d and K_d from the T5, will be applied in section 3.4 for infusing sensory knowledge into T5.

3.4. Sensory attention mechanism in SensoryT5

The sensory vector $s'(w)$ generated by the sensory vector transformation is used as the queries in the attention mechanism of the sensory adapter, substituting the query vector Q in the T5. The sensory adapter performs the attention calculation as follows:

$$A_d = \mathsf{MultiHead}(s'(w), K_0, V_d) = [a_1, a_2, ..., a_i]W_d \qquad (5)$$

where each head is computed as:

$$a_i = \mathsf{Attention}(s'(w)W_i^Q, K_0 W_i^K, V_d W_i^V)$$
$$= \mathsf{Softmax}\left(\frac{(s'(w)W_i^Q)(K_0 W_i^K)^T}{\sqrt{d_k}}\right) V_d W_i^V \qquad (6)$$

Once the output $A_d = a_1, a_2, ..., a_n$ of the sensory adapter is obtained, we apply dropout and pooling operations to form a final representation P_d, which is then used as the input to the classification layer.

$$P_d = \mathsf{Dropout}(\mathsf{Pool}(A_d)) \qquad (7)$$

The pooled representation P_d is then fed into the classifier of the T5.

$$C_d = \mathsf{Softmax}(Linear(\mathsf{Dropout}(P_d))) \qquad (8)$$

C_d is a probability distribution vector. The class with the highest probability is selected as the predicted label, denoted as y.

The first step of the back-propagation process involves computing the gradient of the loss function with respect to the parameters of sensory attention adapter. Θ_A represents the parameters of the sensory attention layer, and A_d represents the output of the sensory T5. The computed gradient is used to update the parameters of the attention layer, enhancing its capacity to integrate sensory information into the T5 model. This is computed as follows:

$$\frac{\partial \mathcal{L}}{\partial \Theta_A} = \frac{\partial \mathcal{L}}{\partial A_d} \cdot \frac{\partial A_d}{\partial \Theta_A} \qquad (9)$$

After the gradients for the sensory attention mechanism have been computed, we then compute the gradients for the parameters of the final layer of the T5, denoted as Θ_E.

$$\frac{\partial \mathcal{L}}{\partial \Theta_E} = \frac{\partial \mathcal{L}}{\partial V_d} \cdot \frac{\partial V_d}{\partial \Theta_E} \quad (10)$$

Finally, the gradients for the sensory information transformation, denoted as Θ_S, are computed as follows:

$$\frac{\partial \mathcal{L}}{\partial \Theta_S} = \frac{\partial \mathcal{L}}{\partial s'(w)} \cdot \frac{\partial s'(w)}{\partial \Theta_S} \quad (11)$$

Here, Θ_S represents the parameters of the sensory information transformation component, which includes the weights and biases of the two linear layers, and $s'(w)$ represents the output of this component. The calculated gradient is used to update the parameters of the sensory information transformation to improve its ability to capture and model sensory information. Through these calculations, we are able to update the parameters of the sensory attention mechanism, the T5, and the sensory information transformation component.

4. Experimental evaluation

4.1. Datasets

We have selected four benchmark datasets of varying sizes to encompass a variety of classification tasks: Empathetic Dialogues (ED) (Rashkin et al., 2019), GoEmotions (GE) (Demszky et al., 2020a), ISEAR (Scherer and Wallbott, 1994) and EmoInt (Mohammad and Bravo-Marquez, 2017). For the GE dataset, we exclusively utilize samples with a single label and omit those that are neutral to maintain an equitable comparison with prior studies (Suresh and Ong, 2021a; Chen et al., 2023). Table 1 presents a summary of key statistics for these datasets. Our evaluation utilizes two widely recognized performance metrics: accuracy and the F1 score, in line with state-of-the-art studies.

Dataset	N_{train}	N_{test}	L	C
ED	19,533	2,547	18	32
GE	23,485	2,984	12	27
ISEAR	4,599	1,534	22	7
EmoInt	3,612	3,141	16	4

Table 1: Statistics of the four benchmark datasets. In the table, "N_{train}" and "N_{test}" respectively represent the number of instances in the training and testing sets. "L" stands for the average text length within the dataset, and "C" indicates the number of classes/categories.

4.2. Sensory knowledge

Before conducting the emotion analysis experiments, we conducted a preliminary analysis of our sensory lexicons from the perspective of sensory perception value distribution. Figure 2 displays histograms of the six sensory measures across all words within our model. Notably, the distributions for these measures are quite unbalanced. Gustatory and olfactory measures predominantly demonstrate a left-skewed distribution, with most values ranging between 0 and 1. This suggests that these two sensory perceptions are less frequently represented in the textual context. Thus, it might be challenging to represent gustatory and olfactory perceptions from text.

In contrast, auditory and visual measures show a relatively uniform distribution. The auditory measure is evenly distributed between 0 and 2.5, while the visual measure ranges between 2 and 4.5. These distributions indicate a higher sensitivity of auditory and visual knowledge to textual information, which suggests that auditory and visual senses may play a significant role within sensory models.

Lastly, haptic and interoceptive measures exhibit similar trends, declining from about 2500 to 0 as the values increase from 0 to 5. The decline in the presence of haptic and interoceptive knowledge across the general textual context might suggest that they are less informative sensory dimensions in the majority of cases.

As discussed in section 3.1, the Lancaster Sensorimotor Norms dataset is subject to size limitations, resulting in a significant number of unknown words for which corresponding sensory values are unavailable. To address this challenge, we adopted the method proposed by Li et al. (2017) for predicting sensory values of unknown words through embedding techniques. In our experiments, we utilized both the T5 embedding and the GloVe embedding (Pennington et al., 2014b) for this prediction task.

To assess the accuracy of our predictions, we randomly selected 10% of the Lancaster Sensorimotor Norms dataset as a validation set and applied the Root Mean Square Error (RMSE) as the evaluation metric. The experimental results are presented in Table 2. The results demonstrate that GloVe outperforms T5 Embedding in predicting each sensory dimension. To preserve the original features of the Lancaster dataset to a minimal extent, we opted for a smaller version of GloVe with 400,000 data points and 200 dimensions. Following augmentation, our sensory vocabulary size reached 407,572[1].

For validating our augmentation, we evaluated the coverage rates of sensory word vectors before and after augmentation across all datasets we employed, as detailed in Table 3. As evident from the augmentation results, the coverage range significantly expands in comparison to the original data across all datasets. This underscores the en-

---

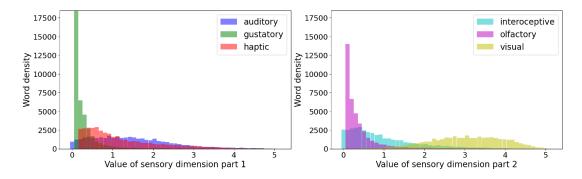[1]We will release the sensory vector after this paper is accepted

Figure 2: Histograms showing the distribution six sensory values over words. X-axis shows the value in an sensory dimension, while y-axis displays the word density.

Sensory Name	T5 Embedding	GloVe
Auditory	0.949	**0.803**
Gustatory	0.632	**0.534**
Haptic	0.893	**0.698**
Interoceptive	0.831	**0.662**
Olfactory	0.572	**0.501**
Visual	0.842	**0.743**
Total	0.798	**0.665**

Table 2: Comparison of prediction accuracy between T5 Embedding and GloVe techniques on different sensory dimensions, as measured by RMSE values. Lower scores indicate higher accuracy in the prediction of sensory values.

hanced impact of integrating sensory information into the model on the results.

Datasets	Lancaster %	Exten-Lancaster %
ED	58.23	91.78
GE	46.85	83.91
ISEAR	54.62	78.97
EmoInt	29.65	46.21

Table 3: Word coverage of Lancaster Sensorimotor Norms before and after expansion using regression prediction.

4.3. Experiment settings and Baselines

We compare the proposed SensoryT5 primarily with two group of strong baselines:

PLMs. We compared against BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), XLNet (Yang et al., 2019) and T5 (Raffel et al., 2020). The advent of PLMs has marked a significant improvement across a multitude of tasks in the realm of natural language processing, including text classification. This leap in performance is largely due to the deep and nuanced semantic representations these models extract from the text, facilitating a more profound understanding and interpretation of linguistic content.

Label Embedding-aware models. Suresh and Ong (2021a) introduced a concept called label-aware contrastive loss (LCL). This technique uniquely assigns varying weights to each negative sample. Importantly, pairs that are more easily confounded have a higher impact on the objective function, enhancing outcomes in fine-grained text classification scenarios. Chen et al. (2023) proposed HypEmo, a framework enhancing fine-grained emotion classification by utilizing hyperbolic space for label embedding. This model integrates hyperbolic and Euclidean geometries to discern subtle nuances among labels effectively.

These two models, LCL and HypEmo, stand as the most potent in the realm of fine-grained emotion classification, delivering unparalleled results due to their innovative handling of nuanced label distinctions and hierarchical intricacies.

Implementation Details.[2] During training, we applied the Adam optimizer in Euclidean space. We set the learning rate at a consistent 10^{-4}, maintaining a balance between rapid adaptation and the stability of learning, reducing the likelihood of oscillation or divergence.

4.4. Baseline comparison

To demonstrate the effectiveness of SensoryT5, we embarked on a comprehensive set of comparative experiments, analyzing its performance in emotion classification tasks. The comparison is shown in Table 4. Firstly, we compare SensoryT5 with PLMs. SensoryT5 registers an impressive enhancement over T5's performance, the best of the PLM contenders. For instance, SensoryT5 exhibits an increase in accuracy by 0.9% for Empathetic Dialogues and 1.3% for GoEmotions, showcasing its finesse in handling diverse emotional contexts. This upward trend continues with ISEAR and EmoInt datasets, where SensoryT5 improves by 0.9% and 1.2%, respectively, over T5.

---

[2]We will release the open-source code after this paper is accepted

	Empathetic Dialogue		GoEmotions		ISEAR		EmoInt	
	ACC	F1	ACC	F1	ACC	F1	ACC	F1
⋆BERT$_{large}$	0.557	0.551	0.642	0.637	0.677	0.679	0.848	0.848
⋆RoBERTa$_{large}$	0.596	0.590	0.652	0.644	0.723	0.720	0.865	0.865
⋆XLNet$_{large}$	0.599	0.592	0.641	0.568	0.711	0.711	0.845	0.845
⋆T5$_{large}$	0.609	0.604	0.661	0.657	0.717	0.717	0.863	0.863
†LCL	0.601	0.591	0.655	0.648	0.724	0.724	0.866	0.866
§HypEmo	0.596	0.610	0.654	0.663	0.707	0.712	0.846	0.846
⋆SensoryT5	**0.618**	**0.615**	**0.674**	**0.670**	**0.726**	**0.724**	**0.875**	**0.875**

Table 4: Evaluation on fine-grained emotion classification, the result with the best performance are highlighted in bold. Data marked with †are from (Suresh and Ong, 2021b), §from (Chen et al., 2023), and ⋆represents our own results. Note: In §, results from missing datasets (ISEAR and EmoInt) were supplemented by our experiments.

Secondly, we compare with label-aware system. These two models, LCL and HypEmo, stand as the most potent in the realm of fine-grained emotion classification. LCL outperforms T5 in the ISEAR and EmoInt datasets, while the other datasets under the label-aware system category do not compete favorably with T5. This comparative analysis is critical, considering that LCL utilizes a synonym substitution technique to effectively double its dataset size. Such an expansion contributes significantly to its enhanced performance metrics. In our experiments, we strictly adhered to using original samples without resorting to any form of data augmentation techniques. Despite this, SensoryT5 surpasses LCL by 0.2% and 0.9% in accuracy on the ISEAR and EmoInt datasets, respectively. This margin of improvement, although seemingly nominal, is quite significant in the context of these tasks. It underscores the efficacy of our proposed method of infusing sensory perceptions into the model.

In summary, compared to previous studies, we have achieved superior results without the necessity for additional data, marking the current pinnacle in this field. This accomplishment underscores the effectiveness of SensoryT5.

4.5. Ablation studies

In our efforts to understand the contributions of different components within the SensoryT5 model, we conducted ablation studies, a critical methodological step in assessing the impact of our novel sensory integration. These studies were also carried out on four datasets. The ablation tests were structured around three primary configurations:

SensoryT5: Our complete model infusing sensory information.

Random SensoryT5: A variant of our model where the sensory values were substituted with random numbers ranging from 0 to 5, maintaining the same distribution of sensory scores but eliminating their meaningful association with the data.

T5 (None): The baseline model without any sensory information, representing the standard PLM approach in fine-grained emotion classification tasks.

The result is shown in Figure 3. While the SensoryT5 model exhibited the highest performance in terms of accuracy across all datasets, the Random SensoryT5 configuration yielded lower results than even the T5. This decrement in performance was especially pronounced on the more complex datasets, Empathetic Dialogues and GoEmotions.

The degradation in performance with random sensory values underscores the importance of meaningful sensory integration. It is not merely the presence of additional numerical data that enhances the SensoryT5 model's performance, but rather the contextually relevant and accurately associated sensory information that it brings to the emotion classification task.

Furthermore, the fact that the Random SensoryT5 underperformed compared to the T5 indicates that arbitrarily added sensory information could introduce noise into the model, disrupting its ability to correctly interpret and classify emotional content. This revelation is significant, affirming that the strategic integration of sensory data is crucial, and haphazard integration could be counterproductive.

In summary, these ablation studies have confirmed the value of our sensory information layer, as evidenced by the performance drop when this layer is randomized or removed. This reinforces our assertion that the SensoryT5's strength lies in its ability to simulate a more human-like understanding of textual data, resonating with how humans perceive emotions through a sensory lens.

4.6. Case study

We conducted a focused case study on the SensoryT5 model using a sentence from the Empathetic Dialogues dataset: "I get so mad when I see or hear about kids getting bullied..." In Figure 4, attention heatmaps display the model's focus during processing. The SensoryT5 heatmap shows the aggregate attention for each token in the sensory layer, while the T5 section compiles attention weights across
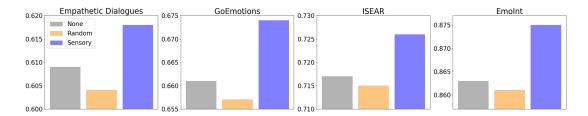
Figure 3: Ablation Study Results. Performance of T5 (None), Random SensoryT5 (with sensory values randomly assigned), and SensoryT5 across four datasets, evaluated using accuracy as the metric.
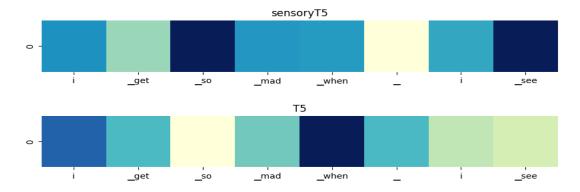

Figure 4: The heat values of the final sensory layer in SensoryT5 and the encoder layer in T5 for the sentence 'I get so mad when I see or hear about kids getting bullied...' sourced from the Empathetic Dialogues training dataset.

all encoder layers, subsequently averaging them to reveal the model's overall focus. The SensoryT5 model exhibited intensified attention on the emotionally significant phrase "so mad," highlighting its ability to detect crucial emotional nuances. In contrast, the standard T5's attention was more distributed, less focused on the emotional pivot. This micro-level analysis reveals SensoryT5's superior capability in recognizing emotional cues. Such insights substantiate the efficacy of integrating sensory awareness into language models for improved emotional discernment.

In summary, our extensive evaluations and comparative studies highlight the superior performance of SensoryT5 over other PLMs based emotion classification models, including the T5. When benchmarked against the state-of-the-art methods, SensoryT5 notably surpassed them, establishing a new standard in the field. Further, our ablation studies convincingly demonstrate that the effectiveness of SensoryT5 is attributed more to its integration of sensory perception than to structural enhancements. This assertion is corroborated by our detailed case studies, which offer a microscopic view into the instances where SensoryT5's unique capabilities are distinctly evident. Collectively, these findings underscore a breakthrough performance of SensoryT5 in the realm of fine-grained emotion classification. Importantly, it signifies a successful adaptation within the shift towards incorporating neuro-cognitive data in NLP, validating the premise that a deeper convergence between sensory data and language modeling leads to a more profound understanding of emotional nuances.

5. Conclusion

In this paper, we propose the SensoryT5 model designed for the fine-grained emotion classification. This framework harnesses sensory knowledge, aiming to boost the prowess of transformers in pinpointing nuanced emotional subtleties. By integrating sensory knowledge into T5 through attention mechanisms, the model concurrently evaluates sensory cues alongside contextual hallmarks. Crucially, SensoryT5 exhibits exceptional adaptability and precision, making it a formidable tool for tasks in Fine-grained Emotion Classification, including configurations like 32-class, 27-class, 7-class, and 4-class delineations. Moreover, SensoryT5 serves as a conduit between sensory perception and emotional understanding, embodying the recent paradigm shift in NLP towards a more neuro-cognitive approach. It acknowledges and capitalizes on the intrinsic relationship between our sensory experiences and our emotional responses, a connection well-documented in neuro-cognitive science but often under-explored in computational fields. By interpreting sensory lexicon through advanced representation learning, SensoryT5 de-

codes the implicit emotional undertones conveyed, mirroring the human ability to associate sensory experiences with specific emotional states. In recognizing the entwined nature of cognition, sensation and emotive expression, SensoryT5 not only contributes to but also encourages the continuation of interdisciplinary research efforts. It stands as testament to the potential of a more nuanced and integrative approach in NLP, where understanding language transcends the boundaries of words and grammar, delving into the very experiences and perceptions that shape human emotionality.

Limitations

In our work, we utilized GloVe and T5 embeddings to predict sensory values for unknown words using a regression method. This approach learns only from static values. To derive static T5 embeddings, we passed all tokens sequentially through the T5 embedding layer, obtaining a static embedding for each token. This process, however, leads to a limitation: it compromises the original dynamic context-embedding capabilities of T5. In T5 embeddings, different embeddings are obtained based on the different contexts. We intended to learn from these transformer embeddings and then predict. Additionally, when compared to current state-of-the-art models in emotion classification, such as the label embedding-aware HypEmo and LCL, SensoryT5 exhibits certain inadequacies, particularly in terms of interpretability. Both HypEmo and LCL not only surpass SensoryT5 in explaining their decision-making processes but also do so with fewer parameters. These models, by leveraging sophisticated label-aware embedding strategies, provide insights into the nuanced relationships and hierarchies among labels, something that SensoryT5, with its reliance on static values, struggles to achieve. This gap highlights a significant area for improvement in SensoryT5, suggesting the need for an advanced approach that maintains the richness of context-sensitive embeddings while enhancing the model's overall interpretability and efficiency.

6. Bibliographical References

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L_1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

Yoshua Bengio, Patrice Simard, and Paolo Frasconi. 1994. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166.

Mukul Bhalla and Dennis R Proffitt. 1999. Visual–motor recalibration in geographical slant perception. *Journal of experimental psychology: Human perception and performance*, 25(4):1076.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Chih-Yao Chen, Tun-Min Hung, Yi-Li Hsu, and Lun-Wei Ku. 2023. Label-aware hyperbolic embeddings for fine-grained emotion classification. *arXiv preprint arXiv:2306.14822*.

I-Hsuan Chen, Qingqing Zhao, Yunfei Long, Qin Lu, and Chu-Ren Huang. 2019. Mandarin chinese modality exclusivity norms. *PloS one*, 14(2):e0211336.

Xin Chen, Zhen Hai, Suge Wang, Deyu Li, Chao Wang, and Huanbo Luan. 2021. Metaphor identification: A contextual inconsistency based neural sequence labeling approach. *Neurocomputing*, 428:268–279.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Chloe Clavel and Zoraida Callejas. 2015. Sentiment analysis: from opinion mining to human-agent interaction. *IEEE Transactions on affective computing*, 7(1):74–93.

Edward Collins, Nikolai Rozanov, and Bingbing Zhang. 2018. Evolutionary data measures: Understanding the difficulty of text classification tasks. *arXiv preprint arXiv:1811.01910*.

Louise Connell, Dermot Lynott, and Briony Banks. 2018. Interoception: the forgotten modality

in perceptual grounding of abstract and concrete concepts. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 373(1752):20170143.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020a. 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.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ruihai Dong, Michael P O'Mahony, Markus Schaal, Kevin McCarthy, and Barry Smyth. 2013. Sentimental product recommendation. In *Proceedings of the 7th ACM Conference on Recommender Systems*, pages 411–414.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

Paul Ekman et al. 1999. Basic emotions. *Handbook of cognition and emotion*, 98(45-60):16.

Lynn Fainsilber and Andrew Ortony. 1987. Metaphorical uses of language in the expression of emotions. *Metaphor and Symbol*, 2(4):239–250.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Minghui Huang, Haoran Xie, Yanghui Rao, Yuwei Liu, Leonard KM Poon, and Fu Lee Wang. 2020. Lexicon-based sentiment convolutional neural networks for online review analysis. *IEEE Transactions on Affective Computing*, 13(3):1337–1348.

Ozan Irsoy and Claire Cardie. 2014. Opinion mining with deep recurrent neural networks. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 720–728.

Xiaotong Jiang, Qingqing Zhao, Yunfei Long, and Zhongqing Wang. 2022. Chinese synesthesia detection: New dataset and models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3877–3887.

Patrik N Juslin and Daniel Västfjäll. 2008. Emotional responses to music: The need to consider underlying mechanisms. *Behavioral and brain sciences*, 31(5):559–575.

Pei Ke, Haozhe Ji, Siyang Liu, Xiaoyan Zhu, and Minlie Huang. 2019. Sentilare: Sentiment-aware language representation learning with linguistic knowledge. *arXiv preprint arXiv:1911.02493*.

Smith K Khare, Victoria Blanes-Vidal, Esmaeil S Nadimi, and U Rajendra Acharya. 2023. Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations. *Information Fusion*, page 102019.

Zoltán Kövecses. 2019. Perception and metaphor. *Perception metaphors*, 19(327):10–1075.

George Lakoff and Mark Johnson. 1980. Metaphors we live by. *University of Chicago, Chicago, IL*.

Sophia Yat Mei Lee. 2018. Figurative language in emotion expressions. In *Chinese Lexical Semantics: 18th Workshop, CLSW 2017, Leshan, China, May 18–20, 2017, Revised Selected Papers 18*, pages 408–419. Springer.

Minglei Li, Qin Lu, Yunfei Long, and Lin Gui. 2017. Inferring affective meanings of words from word embedding. *IEEE Transactions on Affective Computing*, 8(4):443–456.

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. *arXiv preprint arXiv:2111.02194*.

Zijie Lin, Bin Liang, Yunfei Long, Yixue Dang, Min Yang, Min Zhang, and Ruifeng Xu. 2022. Modeling intra- and inter-modal relations: Hierarchical graph contrastive learning for multimodal sentiment analysis. In *Proceedings of the 29th International Conference on Computational Linguistics*,

pages 7124–7135, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

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*.

Yunfei Long, Rong Xiang, Qin Lu, Chu-Ren Huang, and Minglei Li. 2019a. Improving attention model based on cognition grounded data for sentiment analysis. *IEEE transactions on affective computing*, 12(4):900–912.

Yunfei Long et al. 2019b. A study on using personal profiles for a biased reader emotion prediction model.

Qiang Lu, Xia Sun, Yunfei Long, Zhizezhang Gao, Jun Feng, and Tao Sun. 2023. Sentiment analysis: Comprehensive reviews, recent advances, and open challenges. *IEEE Transactions on Neural Networks and Learning Systems*.

Dermot Lynott, Louise Connell, Marc Brysbaert, James Brand, and James Carney. 2020. The lancaster sensorimotor norms: multidimensional measures of perceptual and action strength for 40,000 english words. *Behavior Research Methods*, 52:1271–1291.

Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 142–150.

Donatas Meškelė and Flavius Frasincar. 2020. Aldonar: A hybrid solution for sentence-level aspect-based sentiment analysis using a lexicalized domain ontology and a regularized neural attention model. *Information Processing & Management*, 57(3):102211.

Saif Mohammad and Felipe Bravo-Marquez. 2017. WASSA-2017 shared task on emotion intensity. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 34–49, Copenhagen, Denmark. Association for Computational Linguistics.

Nadine Müller, Arne Nagels, and Christina Kauschke. 2021. Metaphorical expressions originating from human senses: Psycholinguistic and affective norms for german metaphors for internal state terms (mist database). *Behavior Research Methods*, pages 1–13.

Pansy Nandwani and Rupali Verma. 2021. A review on sentiment analysis and emotion detection from text. *Social Network Analysis and Mining*, 11(1):81.

Paula M Niedenthal and Adrienne Wood. 2019. Does emotion influence visual perception? depends on how you look at it. *Cognition and Emotion*, 33(1):77–84.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Bo Pang, Lillian Lee, et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in information retrieval*, 2(1–2):1–135.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014a. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

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

Rosalind W Picard. 2000. *Affective computing*. MIT press.

Dennis R Proffitt, Mukul Bhalla, Rich Gossweiler, and Jonathan Midgett. 1995. Perceiving geographical slant. *Psychonomic bulletin & review*, 2:409–428.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2018. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207*.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new

benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Cedar R Riener, Jeanine K Stefanucci, Dennis R Proffitt, and Gerald Clore. 2011. An effect of mood on the perception of geographical slant. *Cognition and Emotion*, 25(1):174–182.

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

Goran Šimić, Mladenka Tkalčić, Vana Vukić, Damir Mulc, Ena Španić, Marina Šagud, Francisco E Olucha-Bordonau, Mario Vukšić, and Patrick R. Hof. 2021. Understanding emotions: Origins and roles of the amygdala. *Biomolecules*, 11(6):823.

Richard Socher, Jeffrey Pennington, Eric H Huang, Andrew Y Ng, and Christopher D Manning. 2011. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pages 151–161.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Varsha Suresh and Desmond Ong. 2021a. Not all negatives are equal: Label-aware contrastive loss for fine-grained text classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4381–4394, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Varsha Suresh and Desmond C Ong. 2021b. Not all negatives are equal: Label-aware contrastive loss for fine-grained text classification. *arXiv preprint arXiv:2109.05427*.

Duyu Tang, Bing Qin, and Ting Liu. 2015. Document modeling with gated recurrent neural network for sentiment classification. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1422–1432.

Hao Tian, Can Gao, Xinyan Xiao, Hao Liu, Bolei He, Hua Wu, Haifeng Wang, and Feng Wu. 2020. Skep: Sentiment knowledge enhanced pre-training for sentiment analysis. *arXiv preprint arXiv:2005.05635*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Mingyu Wan, Qi Su, Kathleen Ahrens, and Chu-Ren Huang. 2023. Perceptional and actional enrichment for metaphor detection with sensorimotor norms. *Natural Language Engineering*, pages 1–29.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Rong Xiang, Jing Li, Mingyu Wan, Jinghang Gu, Qin Lu, Wenjie Li, and Chu-Ren Huang. 2021. Affective awareness in neural sentiment analysis. *Knowledge-Based Systems*, 226:107137.

Hu Xu, Bing Liu, Lei Shu, and Philip S Yu. 2019. Bert post-training for review reading comprehension and aspect-based sentiment analysis. *arXiv preprint arXiv:1904.02232*.

Takashi Yamamoto. 2008. Central mechanisms of taste: Cognition, emotion and taste-elicited behaviors. *Japanese Dental Science Review*, 44(2):91–99.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pre-training for language understanding. *Advances in neural information processing systems*, 32.

Sruthi Yarkareddy, T Sasikala, and S Santhanalakshmi. 2022. Sentiment analysis of amazon fine food reviews. In *2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, pages 1242–1247. IEEE.

Da Yin, Tao Meng, and Kai-Wei Chang. 2020. Sentibert: A transferable transformer-based architecture for compositional sentiment semantics. *arXiv preprint arXiv:2005.04114*.

Jonathan R Zadra and Gerald L Clore. 2011. Emotion and perception: The role of affective information. *Wiley interdisciplinary reviews: cognitive science*, 2(6):676–685.

Matthew D Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European*

Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13, pages 818–833. Springer.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Jialin Pan, and Lidong Bing. 2023. Sentiment analysis in the era of large language models: A reality check. *arXiv preprint arXiv:2305.15005*.

Bo Zhao, Jiashi Feng, Xiao Wu, and Shuicheng Yan. 2017. A survey on deep learning-based fine-grained object classification and semantic segmentation. *International Journal of Automation and Computing*, 14(2):119–135.