EmoLLMs: A Series of Emotional Large Language Models and Annotation Tools for Comprehensive Affective Analysis

Zhiwei Liu*
The University of Manchester
Manchester, United Kingdom
zhiwei.liu-2@postgrad.manchester.ac.uk

Kailai Yang
The University of Manchester
Manchester, United Kingdom
kailai.yang@postgrad.manchester.ac.uk

Qianqian Xie†
The University of Manchester
Manchester, United Kingdom
xqq.sincere@gmail.com

Tianlin Zhang
The University of Manchester
Manchester, United Kingdom
zhangtianlin668@gmail.com

Sophia Ananiadou
The University of Manchester
Manchester, United Kingdom
Artificial Intelligence Research Center
Tokyo, Japan
sophia.ananiadou@manchester.ac.uk

ABSTRACT

Sentiment analysis and emotion detection are important research topics in natural language processing (NLP) and benefit many downstream tasks. With the widespread application of large language models (LLMs), researchers have started exploring the application of LLMs based on instruction-tuning in the field of sentiment analysis. However, these models only focus on single aspects of affective classification tasks (e.g. sentimental polarity or categorical emotions), and overlook the regression tasks (e.g. sentiment strength or emotion intensity), which leads to poor performance in downstream tasks. The main reason is the lack of comprehensive affective instruction tuning datasets and evaluation benchmarks, which cover various affective classification and regression tasks. Moreover, although emotional information is useful for downstream tasks, existing downstream datasets lack high-quality and comprehensive affective annotations. In this paper, we propose EmoLLMs, the first series of open-sourced instruction-following LLMs for comprehensive affective analysis based on fine-tuning various LLMs with instruction data, the first multi-task affective analysis instruction dataset (AAID) with 234K data samples based on 3 classification tasks and 2 regression tasks to support LLM instruction tuning, and a comprehensive affective evaluation benchmark (AEB) with 8 regression tasks and 6 classification tasks from various sources and domains to test the generalization ability of LLMs. We propose a series of EmoLLMs by fine-tuning LLMs with AAID to solve various affective instruction tasks. We compare our models with a variety of LLMs and sentiment analysis tools on AEB, where our models outperform all other open-sourced LLMs and sentiment analysis tools, and surpass ChatGPT and GPT-4 in most tasks, which shows that the series of EmoLLMs achieve the ChatGPT-level and GPT-4-level generalization capabilities on affective analysis tasks, and demonstrates our models can be used as affective annotation tools. This project is available at https://github.com/lzw108/EmoLLMs/.

CCS CONCEPTS

• **Computing methodologies → Natural language processing**; **Language resources**; • **Information systems → Sentiment analysis**.

KEYWORDS

Sentiment analysis, emotion detection, large language models, affective instruction dataset, affective evaluation benchmark

---

*Corresponding author

†Qianqian is now affiliated with Yale University. The work was done while she was at The University of Manchester.

1 INTRODUCTION

Emotions and sentiments play a crucial role in shaping our lives. Our words and actions serve as indicators of our emotional states [27]. Leveraging natural language processing (NLP) techniques such as Emotion Detection (ED) and Sentiment Analysis (SA), we can delve into the analysis of human interactions, enabling us to comprehend people's emotional responses toward particular subjects [18]. Specifically, SA tasks typically involve predicting the polarity (usually positive, negative, or neutral), along with the strength of this tone [41], and emotion detection tasks often involve classifying data into fine-grained emotion categories (e.g. Ekman [16], Plutchik [35]) or predicting the intensity of emotions [46]. These affective information are proven as useful features for many downstream tasks, including mental health analysis [55], misinformation detection [29], and empathetic dialogue systems [31].

Zhiwei Liu, Kailai Yang, Qianqian Xie, Tianlin Zhang, & Sophia Ananiadou

Pre-trained language models (PLMs) such as BERT [11] and RoBERTa [28], have shown excellent performance in classification tasks. Many studies have applied them to sentiment analysis or emotion detection tasks [4, 26, 50]. However, these PLMs are limited by the scale of model parameters and the training corpus, resulting in a lack of comprehensive understanding and generalization ability for complex tasks [52], which restricts the effectiveness of affective analysis, especially in affective regression tasks [53]. Compared to PLMs, LLMs have the characteristic of having an enormous parameter size, typically reaching hundreds of billions or more, which gives them stronger generalization abilities in downstream tasks and enables them to handle tasks with intricate complexities [57]. Many researchers have started exploring the application of LLMs in the field of sentiment analysis, which achieved excellent performance by fine-tuning open-source LLMs on sentiment analysis tasks [24, 52, 57]. However, these studies only focus on sentiment/emotion classification tasks and overlook regression tasks (e.g. sentiment strength, emotion intensity), which provide more fine-grained affective features [1] and are proven useful in many scenarios [9, 36, 56]. The major reason is the lack of a comprehensive instruction-based sentiment analysis dataset and evaluation benchmark.

Moreover, though emotional information is proven useful for downstream tasks, existing downstream datasets lack emotion-related resources such as sentiment/emotion labels. Therefore, many works use affective analysis tools (e.g. VADER [21], TextBlob[1]) to provide sentiment annotations. For example, in [17], the authors utilized the TextBlob library to calculate sentiment scores and fed them into a depression detection classifier. Additionally, some studies employ transfer learning methods, by applying models trained on other sentiment analysis or emotion-labeled datasets to automatically annotate the emotions expressed in downstream task datasets [7, 13]. However, these tools or methods can only annotate one aspect of sentiment analysis tasks, resulting in limited coverage of emotional features.

To address the above issues, we propose a suite of LLMs, instruction-tuning datasets, and an evaluation benchmark for multi-task affective analysis. We first construct the multi-task affective analysis instruction dataset (AAID) with 234K data samples to support LLM instruction tuning, which is based on SemEval-2018 Task1: Affect in Tweet [32, 33], including five tasks: emotion intensity regression, ordinal classification of emotion intensity, sentiment strength regression, sentiment classification, and multi-label emotion classification. Based on the AAID dataset, we propose a series of emotional large language models (EmoLLMs), the first open-sourced instruction-following LLMs for comprehensive affective analysis, by performing multi-task instruction tuning on LLMs. To evaluate the performance and generalizability of EmoLLMs, we also construct an affective evaluation benchmark (AEB) based on 14 affective analysis datasets collected from various platforms and sources, which include 8 regression tasks and 6 classification tasks.

Based on AEB, we evaluate EmoLLMs, a variety of open-sourced LLMs, close-sourced LLMs (i.e. ChatGPT and GPT-4), and several sentiment analysis tools. The experimental results indicate that the series of EmoLLMs overtake all other open-sourced LLMs, and

sentiment analysis tools, and exceed ChatGPT and GPT-4 in 7 regression tasks and 4 classification tasks. These results demonstrate that EmoLLMs achieve a comparable capability with ChatGPT and GPT-4 in most affective analysis tasks. EmoLLMs can serve as comprehensive affective annotation tools for annotating data from different platforms and sources.

Our main contributions are as follows:

- We build AAID, the first multi-task affective analysis instruction tuning data, and AEB, the first affective generalization testing instruction benchmark.
- We introduce a series of EmoLLMs, the first open-source instruction following LLMs for comprehensive affective analysis.
- We compare EmoLLMs with other LLMs on AEB. Additionally, we conduct a comprehensive analysis of the affective analysis capabilities of ChatGPT and GPT-4. Our models achieve SOTA performance on the AEB dataset compared to other open-sourced LLMs and present ChatGPT-level and GPT-4-level generalization capabilities, establishing their potential as effective tools for affective annotation.

The structure of this paper is as follows: Section 2 introduces the related work about sentiment analysis models and open-sourced LLMs. Section 3 introduces the proposed method. Specifically, Section 3.1 introduce the task definition. Section 3.2 and Section 3.3 present the construction process of AAID and AEB respectively. Section 3.4 introduces the training process of EmoLLMs. Section 4 presents the experiment results on AEB and analyses the performance of each model. Section 5 concludes this paper by summarizing our findings. Section 6 discusses the real-world applications of EmoLLMs, limitations, and future work.

2 RELATED WORK

2.1 Affective Analysis Model

There have been various affective analysis tools proposed, such as VADER [21], and TextBlob. Although these tools are convenient to use, their effectiveness in sentiment analysis is not ideal [19]. In recent years, many studies have focused on fine-tuning PLMs to enhance their capabilities in the field of sentiment analysis. Bello et al. [4] combine BERT with other deep learning models (e.g. CNN, RNN, LSTM) to improve the ability of the model in short and simple text sentiment analysis. Liao et al. [26] propose a multi-task model based on RoBERTa for aspect-category sentiment analysis. Yin et al. [50] propose the SentiBERT model, which focuses on the field of sentiment analysis. SentiBERT integrates a recursive constituency tree based on BERT to better capture compositional sentiment semantics. Recently, numerous studies have embarked on investigating the utilization of LLMs in sentiment analysis, resulting in remarkable performance gains in sentiment analysis tasks. Zhang et al. [52] propose a retrieval-augmented LLM for financial sentiment analysis, which utilizes additional background information from external sources and outperforms LLM baselines by 15% and 48%. Similarly, Lei et al. [24] also use a simple yet effective retrieval module to enhance the emotion recognition capability of LLM in dialogue. Zhang et al. [57] develop a context and emotion knowledge-tuned LLM, namely DialogueLLM, obtained by fine-tuning LLM with multimodal (i.e., texts and videos) emotional dialogues, which achieved

SOTA results on three emotion recognition in conversations (ERC) datasets. However, these PLMs and LLMs only focus on individual aspects of affective analysis, lacking the ability to predict sentiment strength and emotion intensity.

2.2 Open Sourced Large Language Models

Although ChatGPT and GPT-4 have shown excellent performance in various fields, their closed-source availability affects the progress of scientific research. Therefore, numerous studies are dedicated to democratizing LLMs, such as the LLaMA series [42, 43], OPT series [54], BLOOM series [45], and Falcon [34]. Based on the open-source LLMs, many efforts have been made to develop models with instruction-following capabilities like ChatGPT by training on extensive instruction-tuning datasets (e.g. Alpaca[2] and the Vicuna[3]). Recently, there has been a lot of domain-specific work aimed at improving the performance of LLM in specific domains by training on domain-specific instruction datasets. Such as FinMA [47] in the finance domain, MentalLLaMA [48] in the mental health domain, TimeLlaMA [51] used for temporal reasoning, and ExTES-LLaMA [58] in emotional support chatbots. Our work is the first open-sourced LLM series for comprehensive multitask affective analysis.

3 METHODS

The goal of this work is to evaluate and enhance the comprehensive and complex affective analysis capabilities of LLMs. To achieve this objective, we build the first affective analysis instruction dataset (AAID) to support LLMs tuning for comprehensive affective analysis tasks. We propose EmoLLMs, a series of emotional LLMs by fine-tuning LLMs based on AAID. Furthermore, we construct a comprehensive affective evaluation benchmark to test the generalization ability of LLMs.

3.1 Task Definition

Similar to [48] in handling mental health analysis tasks, we also approach affective analysis as a generative task, where a generative model (i.e., an autoregressive language model $P_\phi(y|x)$ parameterized by pre-trained weights ϕ) is employed as the foundation, which is unlike previous discriminative and regression models. This model is capable of simultaneously addressing N affective analysis tasks, such as sentiment polarity and strength prediction, emotion classification and intensity prediction. Each task t is represented by a subset of training context-target pairs: $D_t = (q_i^t, r_i^t)_{i=1,2,...N_t}$, where q is a token sequence containing the task description, target text, and query, and r is another sequence containing the query answer (i.e., classification result or regression result). All subsets are combined into a training dataset: D. The model is optimized based on this merged data, aiming to maximize the conditional language modeling objective to enhance the accuracy of predictions.

3.2 Instruction Tuning Data Building

We build the instruction dataset based on the SemEval-2018 Task 1: Affect in Tweets, which includes a series of highly annotated sentiment analysis subtasks [32, 33].

---

[2]https://crfm.stanford.edu/2023/03/13/alpaca.html
[3]https://lmsys.org/blog/2023-03-30-vicuna/

Table 1: Statistics of the data. 'Raw' denotes the raw data from SemEval-2018 Task 1: Affect in Tweets. 'Instruction' denotes the converted instruction data based on raw data.

Task	Raw (Train/Dev)	Instruction (Train/Dev)	Source
EI-reg, EI-oc			
anger	1701/388	17010/3880	Twitter
fear	2252/389	22520/3890	Twitter
joy	1616/290	16160/2900	Twitter
sadness	1533/397	15330/3970	Twitter
V-reg, V-oc	1181/449	11810/4490	Twitter
E-c	6838/886	68380/8860	Twitter

3.2.1 Raw Data. SemEval 2018 Task1 contains five subtasks: 1. emotion intensity regression (**EI-reg**), 2. ordinal classification of emotion intensity (**EI-oc**), 3. valence (sentiment) regression (**V-reg**), 4. ordinal classification of valence (sentiment) (**V-oc**), and 5. emotion classification (**E-c**).

EI-reg: Given a tweet and an emotion E (anger, fear, joy, sadness), determine the intensity of E that best represents the mental state of the tweeter—a real-valued score between 0 (least E) and 1 (most E);

EI-oc: Given a tweet and an emotion E (anger, fear, joy, sadness), classify the tweet into one of four ordinal classes (0: no E can be inferred. 1: low amount of E can be inferred. 2: moderate amount of E can be inferred. 3: high amount of E can be inferred) of intensity of E that best represents the mental state of the tweeter;

V-reg: Given a tweet, determine the intensity of sentiment or valence (V) that best represents the mental state of the tweeter—a real-valued score between 0 (most negative) and 1 (most positive);

V-oc: Given a tweet, classify it into one of seven ordinal classes (from -3: very negative to 3: very positive), corresponding to various levels of positive and negative sentiment intensity, that best represents the mental state of the tweeter;

E-c: Given a tweet, classify it as 'neutral or no emotion' or as one, or more, of eleven given emotions (anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise, trust) that best represent the mental state of the tweeter.

3.2.2 AAID: Affective Analysis Instruction Dataset. We construct the instruction dataset based on the raw data. Due to the limited quantity of the original dataset, we utilize 10 different task instructions for each task to augment the training set and validation set. The data statistics are presented in Table 1. Specifically, we build instruction-tuning samples based on some templates. Table 2 describes the specific instruction templates for each task, and Figure 1 provides corresponding examples (Taking EmoLLaMA as the example and each task selects one [*task prompt*] as an example). [*task prompt*] describes the instructions for each specific task. The word 'Tweet' can be adjusted based on the actual task. The [*input text*] refers to the content of the raw data. The final [*output*] should be adjusted based on the specific task to provide sentiment classification, sentiment strength, emotion classification, or emotion intensity.
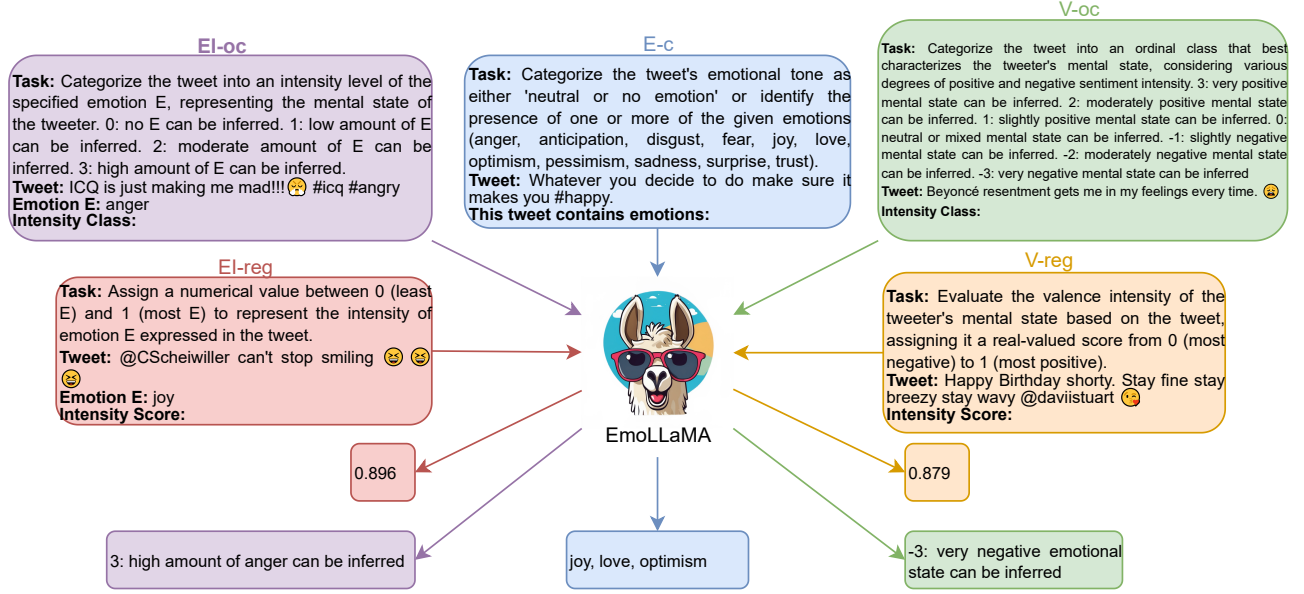
Figure 1: An overview of multi-task instruction tuning of EmoLLaMA for multiple affective analysis tasks.

Table 2: Templates for constructing prompts for instruction dataset. [*task prompt*] denotes task instruction, [*input text*] is from the raw data, [*emotion*] can be anger, fear, joy, sadness, [*output*] is the output from LLM.

Task	Prompt Template
EI-reg	Task: [*task prompt*] Tweet: [*input text*] Emotion E: [*emotion*] Intensity score: [*output*]
EI-oc	Task: [*task prompt*] Tweet: [*input text*] Emotion E: [*emotion*] Intensity class: [*output*]
V-reg	Task: [*task prompt*] Tweet: [*input text*] Intensity score: [*output*]
V-oc	Task: [*task prompt*] Tweet: [*input text*] Intensity class: [*output*]
E-c	Task: [*task prompt*] Tweet: [*input text*] This tweet contains emotions: [*output*]

3.3 AEB: Affective Evaluation Benchmark Building

We first collect the test data from SemEval-2018 Task 1: Affect in Tweets. To test the robustness of our model, a random instruction from the ten instructions used in train augment is selected for each instance in the test set. We also collect additional sentiment analysis or emotion detection datasets from various sources and domains to test the generalizability of our model. We construct the AEB following the template format provided in Table 2. Table 4 shows the task prompt example for each dataset. Except for the four datasets from VADER, all other datasets utilize the original test dataset. Table 3 shows the statistic details.

Datasets used in Valence Aware Dictionary for sEntiment Reasoning (VADER) [22]: There are four datasets from different social media platforms with sentiment intensity (Valence) scores within [-4,4]: **V-Amazon** (Amazon reviews snippets), **V-Movies** (Movies reviews snippets, collected from rotten.tomatoes.com), **V-NYT** (New York Times editorial snippets), **V-Tweet** (Tweets). We randomly sampled 1000 instances from each dataset for generalizability testing.

EmoBank [5, 6]: This dataset was collected from News, blogs, fictions, letters etc. and contains three dimensions, which were manually annotated with emotion according to the psychological Valence-Arousal-Dominance scheme with scores within [1,5].

Stanford Sentiment Treebank (SST) [39]: It is collected from movie reviews, which is the first corpus with fully labeled parse trees, allowing for a comprehensive analysis of the compositionality of sentiment in language. In SST[4], each sentence is assigned a floating-point label that indicates the degree of positive sentiment, ranging from 0.0 to 1.0. while in SST5[5], each sentence is annotated with five labels: very positive, positive, neutral, negative, very negative.

Target Dependent Twitter Sentiment Classification (TDT) [14]: It is a Twitter sentiment classification dataset collected from post comments for the celebrities, products, and companies, which is annotated manually with three labels (negative, neutral, positive). To facilitate our generalizability testing, we restored the entities that were masked in the original data, creating a standard sentence-level sentiment analysis dataset.

GoEmotion [10]: It is a multi-label classification dataset collected from Reddit comments, which consists of 28 emotion labels, including the neutral. However, the original dataset with 28 emotion labels is imbalanced. To mitigate this issue, we select the "Ekman"

---

[4]https://huggingface.co/datasets/sst
[5]https://huggingface.co/datasets/SetFit/sst5

Table 3: Statistics of AEB. 'R' denotes the regression task, followed by the intensity range. 'SC' denotes the sentiment classification task and 'EC' denotes the emotion classification task, followed by the number of categories.

Dataset	Size	Type	Source	Dataset	Size	Type	Source	Dataset	Size	Type	Source
EI-reg	4068	R[0,1]	Twitter	V-Amazon	1000	R [-4,4]	Amazon	SST	2210	R [0,1]	Movie reviews
EI-oc	4068	EC(4)	Twitter	V-Movies	1000	R [-4,4]	Movies reviews	SST-5	2210	SC (5)	Movie reviews
V-reg	937	R[0,1]	Twitter	V-NYT	1000	R [-4,4]	New York Times	TDT	692	SC (3)	Twitter
V-oc	937	SC(7)	Twitter	V-Tweet	1000	R [-4,4]	Twitter	GoEmotion	5427	EC (7)	Reddit
E-c	3259	EC(11)	Twitter	EmoBank	1000	R [1,5]	News, blogs. etc.				

Table 4: The task prompt example for each dataset in AEB.

Dataset	Task prompt
EI-reg	Assign a numerical value between 0 (least E) and 1 (most E) to represent the intensity of emotion E expressed in the tweet.
EI-oc	Categorize the tweet into an intensity level of the specified emotion E, representing the mental state of the tweeter. 0: no E can be inferred. 1: low amount of E can be inferred. 2: moderate amount of E can be inferred. 3: high amount of E can be inferred.
V-reg	Evaluate the valence intensity of the tweeter's mental state based on the tweet, assigning it a real-valued score from 0 (most negative) to 1 (most positive).
V-oc	Categorize the tweet into an ordinal class that best characterizes the tweeter's mental state, considering various degrees of positive and negative sentiment intensity. 3: very positive mental state can be inferred. 2: moderately positive mental state can be inferred. 1: slightly positive mental state can be inferred. 0: neutral or mixed mental state can be inferred. -1: slightly negative mental state can be inferred. -2: moderately negative mental state can be inferred. -3: very negative mental state can be inferred.
E-c	Categorize the tweet's emotional tone as either 'neutral or no emotion' or identify the presence of one or more of the given emotions (anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise, trust).
V-A, V-M, V-NYT, V-T	Calculate the sentiment intensity or valence score of the text, which should be a real number between -4 (extremely negative) and 4 (extremely positive).
SST	Calculate the sentiment score of the text, which should be a real number between 0 (extremely negative) and 1 (extremely positive).
Emobank	Determine the valence/arousal/dominance intensity of the writer's mental state on a scale of 1 (most negative) to 5 (most positive).
GoEmotion	Categorize the text's emotional expression, classifying it as either 'neutral' or as one or more of the specified emotions (anger, disgust, fear, joy, sadness, surprise) that reflect the writer's state of mind.
SST5	Classify the text into one of five classes of sentiment that best represents the mental state of the text. 0: very negative, 1: negative, 2: neutral, 3: positive, 4: very positive.
TDT	Classify the text into one of three classes of sentiment that best represents the mental state of the text. -1: negative, 0: neutral, 1: positive.

option from the dataset provided by the authors, which consists of 7 emotion labels, including the neutral.

Since the first five datasets are collected from the same sources as the AAID, the remaining data comes from different platforms and sources, we divide AEB into two parts for comparison. The former is referred to as AEB-1, used to test the training effectiveness of the models. The latter is called AEB-2, which is suitable for testing the generalization ability of models.

3.4 EmoLLMs

We build EmoLLMs by fine-tuning various LLMs based on AAID. We train three EmoLLaMA models based on LLaMA2 [43]: EmoLLaMA-7B, EmoLLaMA-chat-7B, EmoLLaMA-chat-13B by fine-tuning LLaMA2-7B, LLaMA2-chat-7B, LLaMA2-chat-13B, where LLaMA2-chat-7B and LLaMA2-chat-13B are the first open-source LLMs tuned with reinforcement learning from human feedback (RLHF) [40]. We also train EmoOPT, and EmoBLOOM based on OPT-13B [54] and BLOOM-7B [45]) respectively. All models are trained for three epochs based on AdamW optimizer [30], utilizing early stopping techniques [12] to prevent overfitting, and leveraging DeepSpeed [38, 49] to reduce memory usage. We set the batch size to 256. The initial learning rate is set to 1e-6 with a warm-up ratio of 5%, and the maximum model input length is set to 2048. All models are trained on two Nvidia Tesla A100 GPUs, each with 80GB of memory.

4 EVALUATION

4.1 Base Models

PLMs: Sentiment analysis and emotion detection are typically regarded as classification tasks, while intensity prediction is considered a regression task. We select some commonly used PLMs as baseline models, which can only fine-tuned on a single task, including BERT, RoBERTa, and one domain-specific pre-trained model

(i.e. SentiBERT [50]). We add a fully connected neural layer to each model, which is used for classification or regression. For EI-reg and V-reg tasks, we utilize the mean squared error (MSE) loss function. For EI-oc and V-oc tasks, we use cross-entropy loss. For multi-label task E-c, we adopt binary cross-entropy with logits loss.

Zero-shot/few-shot methods (LLMs without fine-tuning): With the emergence of LLMs, zero-shot and few-shot learning have become effective approaches for solving numerous tasks. We select Falcon-7b-instruct [34], Vicuna-13b-v1.5[6], LLaMA2-chat-7B and LLaMA2-chat-13B to perform zero-shot prompting on the instruction dataset. In addition, we employ zero-shot and few-shot prompting methodologies with the closed-source LLM ChatGPT (gpt-3.5-turbo) and GPT-4 (gpt-4-1106-preview). We select at least one piece of data for each emotion category or label category to serve as few-shot prompts.

Emotion-based instruction-tuning methods: In addition to the EmoLLMs series models, we also fine-tuned BART [25], T5 [37] using the same instructional dataset as baseline models to further evaluate the effectiveness of our models.

4.2 Evaluation Methods

For AEB-1, we use the official evaluation metric[7], Pearson correlation coefficient (pcc), as the evaluation metric for EI-reg, EI-oc, V-reg, and V-oc and use accuracy, micro-F1 (mi-F1), macro-F1 (ma-F1) for E-c. Additionally, the official evaluation also incorporates secondary evaluation metrics. For the regression tasks, they also use pearson correlation for a subset of the test set that includes only those tweets with intensity score greater or equal to 0.5. For the ordinal classification tasks, they also use pearson correlation for a subset of the test set that includes only those tweets with

---

[6]https://huggingface.co/lmsys/vicuna-13b-v1.5
[7]https://competitions.codalab.org/competitions/17751

intensity classes low X, moderate X, or high X (where X is an emotion), and use weighted quadratic kappa on the full test set, and adopt weighted quadratic kappa on the some-emotion subset of the test set. Due to space limitations, we do not list the secondary evaluation, and its conclusions are consistent with the primary evaluation results.

For AEB-2, we apply accuracy, and macro-F1 as evaluation metrics for affective classification tasks. For regression tasks, we use the Pearson correlation coefficient (pcc) [8] as the evaluation metric.

4.3 Results

4.3.1 Results on AEB-1.

The evaluation results on AEB-1 are shown in Table 5 (The results of open-sourced models are the average of five runs). The first line is the score of the top 1 on the SemEval-2018 Task1 leaderboard.

Comparison between EmoLLMs and PLMs, Zero-shot/few-shot methods: Figure 2 presents the results on AEB-1 of several different kind of methods. For EmoLLMs, we chose EmoLLaMA-chat-13B, which shows the best overall performance, to compare with other categories. The results in Table 5 show that EmoLLaMA-chat-13B outperforms all other LLMs and surpasses the top ranking[8] in the first four tasks of AEB-1. For the complex tasks EI-reg and EI-oc, EmoLLaMA-chat-13B shows high improvement compared to top 1, with respective increases of 3.2% (EmoLLaMA:0.831, top1:0.799) and 6.8% (EmoLLaMA:0.763, top1:0.695). For independent raw task fine-tuning methods, although these PLMs are trained on extensive datasets and fine-tuned separately for each task, the results do not surpass the original top 1 scores. The findings demonstrate that general PLMs are more prone to overlooking important information compared to LLMs when dealing with affective regression tasks and fine-grained sentiment classification tasks. For zero-shot/few-shot methods, we can observe that this category of methods performs poorly compared to other fine-tuning approaches, especially in the EI-reg and EI-oc tasks. This indicates that the LLMs without fine-tuning struggle to handle the issue of emotion intensity effectively (We also test BART, T5, OPT and BLOOM in zero-shot and few-shot methods, but their response is highly irrelevant).

Comparison between EmoLLMs: We can observe from Table 5 that EmoLLMs all perform well compared with LLMs without fine-tuning. EmoT5 performs the best on the emotion classification task E-c (ma-F1) (EmoT5:0.568, EmoLLaMA:0.545), but it does not perform as well as other models on regression tasks (e.g. EI-reg(ave): EmoT5:0.783, EmoLLaMA:0.831). Although EmoOPT slightly outperforms EmoLLaMA in a few regression tasks (e.g. V-reg: EmoOPT:0.887, EmoLLaMA:0.886), it still lags behind EmoLLaMA in most tasks.

In conclusion, our proposed instruction-tuning strategy for sentiment analysis tasks outperforms PLMs and all LLMs without fine-tuning, achieving the best comprehensive performance. Compared to other instruction-tuned EmoLLMs, EmoLLaMA demonstrates a more comprehensive and integrated capability in affective analysis.

4.3.2 Results on AEB-2.

In order to evaluate the generalizability of EmoLLMs, we execute experiments on the AEB-2 that are not included in the training process (Detailed descriptions can be found in Table 3). All models apply zero-shot method. We compare the series of EmoLLMs with ChatGPT, GPT4, several open-source LLMs (i.e. LLaMA2-chat, Falcon, and Vicuna) and several sentiment analysis tools (i.e. VADER, TextBlob). Table 6 presents the experiment results (The results of open-sourced models are the average of five runs). For EmoLLMs, it is worth noting that, since we use labels ranging from 0 to 1 when fine-tuning the model on the regression dataset, we also use the range of 0 to 1 for predictions during the generalization testing of regression tasks. Afterward, we map these predictions to the corresponding range of the data.

Comparision between EmoLLMs and LLMs without fine-tuning: Figure 3 presents the results on AEB-2 of several different kind of methods. We still choose EmoLLaMA as the representative for EmoLLMs. From Table 6, we can see that EmoLLaMA series outperform ChatGPT, GPT-4, and LLMs without fine-tuning in most regression tasks. In the first four regression tasks, EmoLLaMA overtakes GPT-4 by over 10%. Although EmoLLaMA performs less well than ChatGPT and GPT-4 in SST and Emobank-Arousal, the difference is less than 5%. For classification tasks, EmoLLaMA series performs better than ChatGPT and GPT-4 in the TDT task. In the GoEmotion, the performance of EmoLLaMA is within a 5% difference compared to ChatGPT and GPT-4. In SST5 tasks, GPT-4 performs exceptionally well (acc:0.543, ma-F1:0.504), as we can see that ChatGPT, GPT-4, both outperform other models in SST5 and SST tasks. The possible reason is that the SST dataset is popular, and LLMs have been exposed to similar corpora during pre-training, which enables them to perform better using zero-shot methods.

Comparision between EmoLLMs: Table 6 shows that all instruction tuning LLMs perform well on AEB-2 and have good transferability except EmoBART and EmoT5. EmoBART and EmoT5 perform similarly to their performance on the AEB-1 dataset, showing poor performance in regression tasks. Interestingly, EmoLLaMA-chat-7B performs the best in most tasks of the AEB-2 and even outperforms EmoLLaMA-chat-13B in most regression tasks. One possible reason is that models with a larger number of parameters tend to overfit during fine-tuning, which can subsequently affect their general performance ability.

It is worth noting that, in AEB-2 dataset, only TDT and V-Tweet are sourced from Twitter, while the others are collected from different platforms and domains. Although EmoLLMs' training data is only sourced from Twitter, it performs well on other platforms and domains, which demonstrates its excellent transferability. The results also show that the performance of the current sentiment analysis tools (i.e. VADER, TextBlob) is significantly inferior to that of EmoLLMs. Overall, the experiment results on AEB-2 illustrate EmoLLMs series achieves ChatGPT-level and GPT-4-level general capabilities (especially EmoLLaMA) and can be used as emotion annotation tools.

4.3.3 Analysis of Chatgpt and GPT-4 .

On the AEB-1 dataset, Table 5 shows that GPT-4 and GPT-4-FS perform best in zero-shot/few-shot methods, followed by ChatGPT and ChatGPT-FS. This illustrates the current open-sourced LLMs still have a big gap with ChatGPT and GPT-4 in complex tasks (e.g. gaps between GPT-4

---

[8]Seernet [15] achieved first position in the first four tasks of SemEval-2018 Task1 during the competition phase. It is based on traditional machine learning methods, which perform comprehensive data pre-processing and apply the stacking technique to ensemble multiple ML methods (e.g. XG Boost, Random Forest).

Table 5: Evaluation results on AEB-1. Some results are referenced from [20, 50, 53]. 'FS' denotes few-shot method. Unmarked LLMs all adopt zero-shot method. 'ave' denotes macro-average. 'acc' denotes accuracy. 'mi-F1' denotes micro-F1. 'ma-F1' denotes macro-F1. The evaluation metric for the first four tasks is the Pearson correlation coefficient. The first line is the score of the top 1 on the SemEval-2018 Task1 leaderboard.

| model	EI-reg					EI-oc					V-reg	V-oc		E-c	
	ave	anger	fear	joy	sadness	ave	anger	fear	joy	sadness	valence	valence	acc	mi-F1	ma-F1
Leaderboard(1)	0.799	0.827	0.779	0.792	0.798	0.695	0.706	0.637	0.720	0.717	0.873	0.856	0.609	0.724	0.592
PLMs															
BERT-base	0.785	0.800	0.781	0.783	0.742	0.683	0.698	0.656	0.712	0.665	0.840	0.805	0.567	0.718	0.568
RoBERTa-base	0.717	0.670	0.736	0.769	0.694	0.664	-	-	-	-	0.845	0.772	0.563	0.721	0.536
SentiBERT	0.722	0.724	0.740	0.731	0.691	0.665	-	-	-	-	0.835	0.763	0.535	0.700	0.522
Zero-shot/few-shot methods															
Falcon	0.114	0.147	0.082	0.095	0.131	0.033	0.022	0.017	0.031	0.061	0.135	0.189	0.190	0.318	0.253
Vicuna	0.281	0.307	0.257	0.260	0.299	0.214	0.238	0.193	0.186	0.241	0.298	0.579	0.220	0.359	0.253
LLaMA2-7B-chat	0.194	0.176	0.257	0.097	0.247	0.120	0.112	0.138	0.115	0.114	0.094	0.497	0.257	0.414	0.286
LLaMA2-13B-chat	0.488	0.524	0.506	0.398	0.526	0.194	0.262	0.178	0.119	0.216	0.312	0.568	0.274	0.424	0.302
ChatGPT	0.599	0.637	0.573	0.569	0.618	0.455	0.500	0.428	0.363	0.529	0.637	0.748	0.382	0.546	0.429
ChatGPT-FS	0.550	0.572	0.482	0.587	0.560	0.473	0.502	0.410	0.407	0.573	0.739	0.791	0.413	0.563	0.466
GPT-4	0.656	0.699	0.575	**0.686**	0.667	**0.620**	0.656	0.579	0.618	0.629	0.811	0.788	0.444	0.572	0.497
GPT-4-FS	**0.679**	0.704	0.654	0.679	**0.678**	0.562	0.623	0.523	0.515	0.585	0.825	0.793	**0.460**	**0.582**	**0.515**
Emotion-based instruction-tuning methods															
EmoBART	0.795	0.798	0.803	0.795	0.782	0.725	0.705	0.742	0.723	0.729	0.851	0.835	0.528	0.686	0.548
EmoT5	0.783	0.785	0.797	0.798	0.751	0.717	0.703	0.733	0.726	0.707	0.852	0.836	**0.559**	**0.712**	**0.568**
EmoOPT	0.825	**0.827**	0.830	0.837	0.805	0.753	0.739	0.751	0.762	**0.759**	**0.887**	0.843	0.532	0.680	0.550
EmoBLOOM	0.791	0.802	0.797	0.790	0.776	0.732	0.725	0.717	0.746	0.740	0.857	0.822	0.528	0.683	0.552
EmoLLaMA-7B	0.822	0.819	0.821	0.837	0.809	0.743	0.738	0.722	0.768	0.745	0.879	0.843	0.545	0.695	0.563
EmoLLaMA-chat-7B	0.824	0.825	0.830	0.832	0.810	0.751	0.748	0.754	0.764	0.739	0.876	0.827	0.534	0.693	0.540
EmoLLaMA-chat-13B	**0.831**	0.827	**0.835**	**0.843**	**0.817**	**0.763**	**0.755**	**0.764**	**0.777**	0.755	0.886	**0.860**	0.537	0.696	0.545
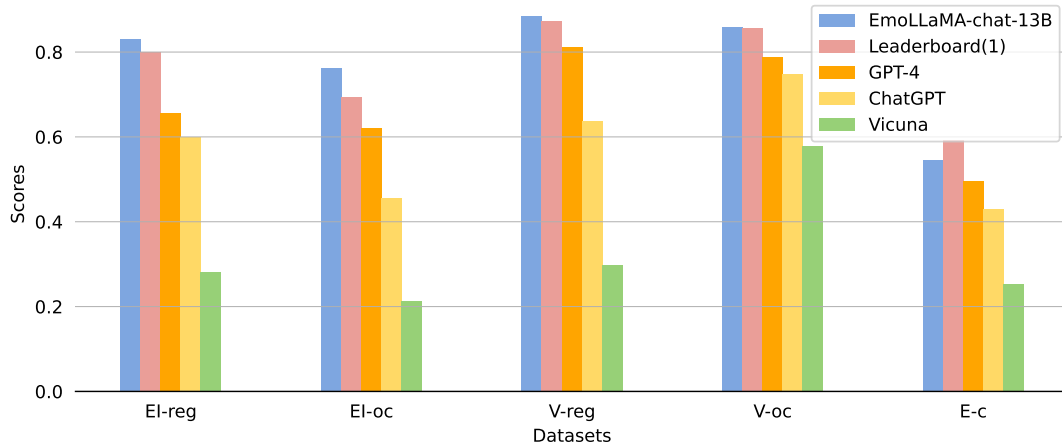
Figure 2: Comparison between EmoLLMs and PLMs, Zero-shot/few- shot methods on AEB-1. The evaluation score for the first four tasks is the pcc (EI-reg and EI-oc adopt macro-average). E-c utilizes macro-F1 score.

and Vicuna: EI-reg(ave): 0.375, EI-oc(ave): 0.426, V-reg: 0.406, V-oc: 0.513, E-c(macro-F1): 0.244). An interesting phenomenon is that in most tasks, ChatGPT and GPT-4 perform better with few-shot method than zero-shot. However, in EI-reg, ChatGPT's zero-shot method outperforms few-shot, while in EI-oc, GPT-4's zero-shot performs better than few-shot. A possible reason is that the EI-reg and EI-oc tasks are more complex, and few-shot learning requires careful design in order to improve the model's performance. For different models, there may be different understandings of few-shot examples. Therefore, for complex tasks, it is necessary to design targeted few-shot examples specifically for different LLMs.

On the AEB-2 dataset, GPT-4 and ChatGPT also perform better than other open-sourced LLMs without fine-tuning in most tasks. However, compared to the performance on the AEB-1 dataset, the performance gap between ChatGPT, GPT-4, and other LLMs without fine-tuning becomes smaller in several tasks (e.g. gaps between GPT-4 and Vicuna: V-A: -0.018, V-M: 0.135, and GoEmotion(acc):0.077). One possible reason is that these tasks are simpler compared to the tasks in AEB-1. This further demonstrates that ChatGPT and GPT-4 are more adept at handling complex tasks compared to other open-source LLMs.

Table 6: Evaluation results on AEB-2 dataset. All evaluation results are based on the zero-shot approach. 'pcc' denotes Pearson correlation coefficient. 'V' denotes Valence. 'V-A' denotes V-Amazon. 'V-M' denotes V-Movies. 'V-T' denotes V-Tweet. 'A' denotes arousal. 'D' denotes dominance. 'acc' denotes accuracy.

| model	V-A	V-M	V-NYT	V-T	SST	EmoBank			GoEmotion		SST5		TDT	
	pcc	pcc	pcc	pcc	pcc	V-pcc	A-pcc	D-pcc	acc	ma-F1	acc	ma-F1	acc	ma-F1
VADER	0.565	0.446	0.464	0.862	0.450	-	-	-	-	-	-	-	0.510	0.266
TextBlob	0.490	0.372	0.288	0.666	0.408	-	-	-	-	-	-	-	0.434	0.435
Falcon	0.492	0.530	0.340	0.449	0.205	0.092	0.059	0.009	0.168	0.223	0.231	0.179	0.355	0.285
Vicuna	0.634	0.592	0.320	0.580	0.733	0.184	0.140	0.002	0.250	0.307	0.293	0.253	0.312	0.269
LLaMA2-chat-13B	0.348	0.479	-0.011	0.300	0.811	0.237	0.248	-0.036	0.278	0.337	0.346	0.281	0.436	0.437
ChatGPT	0.601	0.709	0.419	0.560	0.854	0.554	0.320	-0.121	0.342	**0.407**	0.500	0.397	0.552	0.559
GPT4	0.616	0.727	0.510	0.778	**0.872**	0.723	**0.364**	0.193	0.327	0.401	**0.543**	**0.504**	0.532	0.538
EmoBART	0.770	0.661	0.650	0.853	0.634	0.670	0.059	0.101	0.318	0.366	0.341	0.323	0.529	0.535
EmoT5	0.838	0.728	0.668	0.889	0.724	0.714	0.239	0.064	0.327	0.374	0.373	0.376	0.578	**0.583**
EmoOPT	0.883	0.842	0.767	0.904	0.815	0.713	0.120	**0.247**	0.287	0.331	0.286	0.255	0.573	0.528
EmoBLOOM	0.848	**0.852**	0.649	0.856	0.807	0.663	0.106	0.243	0.302	0.348	0.409	0.353	0.546	0.551
EmoLLaMA-7B	0.866	0.838	0.732	0.902	0.819	0.717	0.219	0.209	0.330	0.366	0.296	0.260	**0.599**	0.579
EmoLLaMA-chat-7B	**0.885**	0.835	**0.797**	**0.910**	0.822	**0.728**	0.192	0.226	**0.371**	0.392	0.400	0.362	0.554	0.554
EmoLLaMA-chat-13B	0.868	0.815	0.780	0.906	0.797	0.726	0.332	0.218	0.350	0.369	0.412	0.399	0.574	0.578
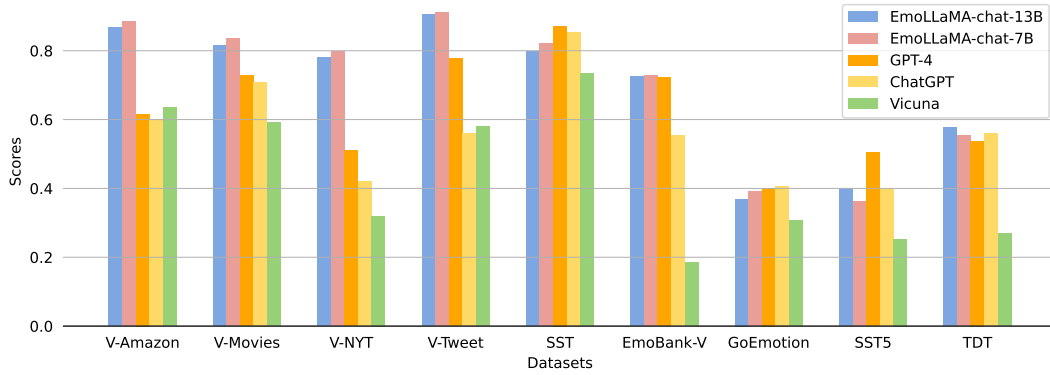
Figure 3: Comparison between EmoLLMs and LLMs without fine-tuning on AEB-2. The evaluation score for the first six tasks (regression tasks) is the pcc. The last three tasks (classification tasks) utilize the macro-F1 score.

To sum up, there is still a certain gap between the current open-source LLMs and ChatGPT, GPT-4 in affective analysis tasks. Currently, we can only surpass ChatGPT and GPT-4 by fine-tuning on specific tasks.

5 CONCLUSION

In this paper, we propose EmoLLMs, a series of comprehensive affective analysis models and annotation tools. We also construct a multi-task affective analysis instruction dataset (AAID) and an affective evaluation benchmark (AEB). We conduct a comprehensive analysis of the performance of EmoLLMs, as well as a variety of LLMs on the AEB benchmark. The results indicate that EmoLLMs perform exceptionally well in both affective analysis regression tasks and classification tasks, achieving SOTA compared to the other open-sourced LLMs, and EmoLLMs exhibit strong transferability, as it has achieved the generalization capabilities of ChatGPT and GPT-4 in various unseen affective analysis tasks. The results also show that there is still a certain gap between the current open-sourced LLMs and ChatGPT, GPT-4 in specific domains. An ideal solution to address the issue is the instruction-tuning strategy employed in

this article, which can greatly enhance the performance of LLMs in a specific domain and surpass ChatGPT and GPT-4 in most tasks.

6 DISCUSSIONS

Real-World Applications. EmoLLMs can provide high-quality and multiple emotional information automatically, which can be used for various practical applications. For example, (1) Misinformation detection: Rumors or fake news often convey specific emotions. Affective features can help verify misinformation [29]. (2) Healthcare (e.g. mental health): The severity of depressive symptoms is closely related to emotions. The main reason is that individuals with depressive symptoms often struggle to regulate their emotions, leading to a decrease in emotional complexity. Therefore, emotional information is useful for diagnosing mental disorders [55]. (3) Customer service (e.g. online shopping): Conducting sentiment analysis on product reviews provides valuable insights into product and service quality as well as customer experience [2].

Limitations and Future Work. Most of the publicly available datasets are from the internet and social media, which have different expression forms, text formats, and styles compared to other types

of textual content. Thus, when applied to the real world, there may be some biases. Additionally, current EmoLLMs are limited to English text content and lack content from other languages and modalities. In the future, we will introduce more datasets from different platforms, domains, modalities, and languages into instruction-tuning data to further enhance the capabilities of EmoLLMs.

ACKNOWLEDGMENTS

REFERENCES

[1] Md Shad Akhtar, Asif Ekbal, and Erik Cambria. 2020. How intense are you? Predicting intensities of emotions and sentiments using stacked ensemble [application notes]. *IEEE Computational Intelligence Magazine* 15, 1 (2020), 64–75.

[2] Hashir Ali, Ehtesham Hashmi, Sule Yayilgan Yildirim, and Sarang Shaikh. 2024. Analyzing amazon products sentiment: a comparative study of machine and deep learning, and transformer-based techniques. *Electronics* 13, 7 (2024), 1305.

[3] BELLEGroup. 2023. BELLE: Be Everyone's Large Language model Engine. https://github.com/LianjiaTech/BELLE.

[4] Abayomi Bello, Sin-Chun Ng, and Man-Fai Leung. 2023. A BERT framework to sentiment analysis of tweets. *Sensors* 23, 1 (2023), 506.

[5] Sven Buechel and Udo Hahn. 2017. Readers vs. writers vs. texts: Coping with different perspectives of text understanding in emotion annotation. In *Proceedings of the 11th linguistic annotation workshop*. 1–12.

[6] Sven Buechel and Udo Hahn. 2022. Emobank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. *arXiv preprint arXiv:2205.01996* (2022).

[7] Jireh Yi-Le Chan, Khean Thye Bea, Steven Mun Hong Leow, Seuk Wai Phoong, and Wai Khuen Cheng. 2023. State of the art: a review of sentiment analysis based on sequential transfer learning. *Artificial Intelligence Review* 56, 1 (2023), 749–780.

[8] Israel Cohen, Yiteng Huang, Jingdong Chen, Jacob Benesty, Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. Pearson correlation coefficient. *Noise reduction in speech processing* (2009), 1–4.

[9] Sergio Consoli, Luca Barbaglia, and Sebastiano Manzan. 2022. Fine-grained, aspect-based sentiment analysis on economic and financial lexicon. *Knowledge-Based Systems* 247 (2022), 108781.

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

[11] 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* (2018).

[12] Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305* (2020).

[13] Diwen Dong, Fuqiang Lin, Guowei Li, and Bo Liu. 2022. Sentiment-Aware Fake News Detection on Social Media with Hypergraph Attention Networks. In *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2174–2180.

[14] Li Dong, Furu Wei, Chuanqi Tan, Duyu Tang, Ming Zhou, and Ke Xu. 2014. Adaptive recursive neural network for target-dependent twitter sentiment classification. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 2: Short papers)*. 49–54.

[15] Venkatesh Duppada, Royal Jain, and Sushant Hiray. 2018. SeerNet at SemEval-2018 Task 1: Domain Adaptation for Affect in Tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*. 18–23.

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

[17] Kuhaneswaran AL Govindasamy and Naveen Palanichamy. 2021. Depression detection using machine learning techniques on twitter data. In *2021 5th international conference on intelligent computing and control systems (ICICCS)*. IEEE, 960–966.

[18] Nida Manzoor Hakak, Mohsin Mohd, Mahira Kirmani, and Mudasir Mohd. 2017. Emotion analysis: A survey. In *2017 International Conference on Computer, Communications and Electronics (Comptelix)*. 397–402. https://doi.org/10.1109/COMPTELIX.2017.8004002

[19] Lu He, Tingjue Yin, and Kai Zheng. 2022. They May Not Work! An evaluation of eleven sentiment analysis tools on seven social media datasets. *Journal of Biomedical Informatics* 132 (2022), 104142.

[20] Amal Htait and Leif Azzopardi. 2021. Sentiment intensity prediction using neural word embeddings. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*. 93–102.

[21] Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, Vol. 8. 216–225.

[22] Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, Vol. 8. 216–225.

[23] Yunjie Ji, Yong Deng, Yan Gong, Yiping Peng, Qiang Niu, Lei Zhang, Baochang Ma, and Xiangang Li. 2023. Exploring the impact of instruction data scaling on large language models: An empirical study on real-world use cases. *arXiv preprint arXiv:2303.14742* (2023).

[24] Shanglin Lei, Guanting Dong, Xiaoping Wang, Keheng Wang, and Sirui Wang. 2023. Instructerc: Reforming emotion recognition in conversation with a retrieval multi-task llms framework. *arXiv preprint arXiv:2309.11911* (2023).

[25] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461* (2019).

[26] Wenxiong Liao, Bi Zeng, Xiuwen Yin, and Pengfei Wei. 2021. An improved aspect-category sentiment analysis model for text sentiment analysis based on RoBERTa. *Applied Intelligence* 51 (2021), 3522–3533.

[27] Bing Liu. 2020. *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge university press.

[28] 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* (2019).

[29] Zhiwei Liu, Tianlin Zhang, Kailai Yang, Paul Thompson, Zeping Yu, and Sophia Ananiadou. 2023. Emotion Detection for Misinformation: A Review. *arXiv preprint arXiv:2311.00671* (2023).

[30] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).

[31] Yukun Ma, Khanh Linh Nguyen, Frank Z Xing, and Erik Cambria. 2020. A survey on empathetic dialogue systems. *Information Fusion* 64 (2020), 50–70.

[32] 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*. 1–17.

[33] 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)*. European Language Resources Association (ELRA), Miyazaki, Japan. https://aclanthology.org/L18-1030

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

[35] Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. In *Theories of emotion*. Elsevier, 3–33.

[36] Syed Arbaaz Qureshi, Gael Dias, Mohammed Hasanuzzaman, and Sriparna Saha. 2020. Improving depression level estimation by concurrently learning emotion intensity. *IEEE Computational Intelligence Magazine* 15, 3 (2020), 47–59.

[37] 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 (2020), 5485–5551.

[38] Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 3505–3506.

[39] 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*. 1631–1642.

---

[9]https://pixlr.com/image-generator/

[40] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems* 33 (2020), 3008–3021.

[41] Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. 2010. Sentiment strength detection in short informal text. *Journal of the American society for information science and technology* 61, 12 (2010), 2544–2558.

[42] 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* (2023).

[43] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).

[44] Cheng Wen, Xianghui Sun, Shuaijiang Zhao, Xiaoquan Fang, Liangyu Chen, and Wei Zou. 2023. ChatHome: Development and Evaluation of a Domain-Specific Language Model for Home Renovation. *arXiv preprint arXiv:2307.15290* (2023).

[45] BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100* (2022).

[46] Hongliang Xie, Shi Feng, Daling Wang, and Yifei Zhang. 2018. A novel attention based CNN model for emotion intensity prediction. In *Natural Language Processing and Chinese Computing: 7th CCF International Conference, NLPCC 2018, Hohhot, China, August 26–30, 2018, Proceedings, Part I 7*. Springer, 365–377.

[47] Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. 2023. PIXIU: A Large Language Model, Instruction Data and Evaluation Benchmark for Finance. *arXiv preprint arXiv:2306.05443* (2023).

[48] Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, and Sophia Ananiadou. 2023. Mentalllama: Interpretable mental health analysis on social media with large language models. *arXiv preprint arXiv:2309.13567* (2023).

[49] Zhewei Yao, Reza Yazdani Aminabadi, Olatunji Ruwase, Samyam Rajbhandari, Xiaoxia Wu, Ammar Ahmad Awan, Jeff Rasley, Minjia Zhang, Conglong Li, Connor Holmes, et al. 2023. DeepSpeed-Chat: Easy, Fast and Affordable RLHF Training of ChatGPT-like Models at All Scales. *arXiv preprint arXiv:2308.01320* (2023).

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

[51] Chenhan Yuan, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2023. Back to the Future: Towards Explainable Temporal Reasoning with Large Language Models. *arXiv preprint arXiv:2310.01074* (2023).

[52] Boyu Zhang, Hongyang Yang, Tianyu Zhou, Muhammad Ali Babar, and Xiao-Yang Liu. 2023. Enhancing financial sentiment analysis via retrieval augmented large language models. In *Proceedings of the Fourth ACM International Conference on AI in Finance*. 349–356.

[53] Linrui Zhang, Hsin-Lun Huang, Yang Yu, and Dan Moldovan. 2020. Affect in Tweets: A Transfer Learning Approach. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. 1511–1516.

[54] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068* (2022).

[55] Tianlin Zhang, Kailai Yang, Shaoxiong Ji, and Sophia Ananiadou. 2023. Emotion fusion for mental illness detection from social media: A survey. *Information Fusion* 92 (2023), 231–246.

[56] Xueyao Zhang, Juan Cao, Xirong Li, Qiang Sheng, Lei Zhong, and Kai Shu. 2021. Mining dual emotion for fake news detection. In *Proceedings of the web conference 2021*. 3465–3476.

[57] Yazhou Zhang, Mengyao Wang, Prayag Tiwari, Qiuchi Li, Benyou Wang, and Jing Qin. 2023. DialogueLLM: Context and Emotion Knowledge-Tuned LLaMA Models for Emotion Recognition in Conversations. *arXiv preprint arXiv:2310.11374* (2023).

[58] Zhonghua Zheng, Lizi Liao, Yang Deng, and Liqiang Nie. 2023. Building emotional support chatbots in the era of llms. *arXiv preprint arXiv:2308.11584* (2023).