IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 1 June 2024, accepted 23 July 2024, date of publication 5 August 2024, date of current version 14 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3438353

RESEARCH ARTICLE

Enhanced Sentiment Intensity Regression Through LoRA Fine-Tuning on Llama 3

DIEFAN LIN, YI WEN, WEISHI WANG, AND YAN SU
Faculty of Arts and Sciences, Beijing Normal University, Zhuhai 519087, China
Corresponding author: Diefan Lin (diefanlin@mail.bnu.edu.cn)

ABSTRACT Sentiment analysis and emotion detection are critical research areas in natural language processing (NLP), offering benefits to numerous downstream tasks. Despite the widespread application of pre-trained models and large language models (LLMs) in sentiment analysis, most previous works have focused on sentiment polarity or emotion classification, neglecting the finer-grained task of sentiment intensity regression, which prevents the precise capture of sentiment intensity and hindering model performance in complex scenarios and diverse applications. To address this issue, we enhance the Roberta model with an efficient additive attention mechanism and an adaptive weighted Huber loss function, notably improving its performance in sentiment intensity regression. Based on the SemEval 2017 and 2018 datasets, we employ prompt engineering to construct fine-tuned datasets, which are further enriched with outputs from the enhanced Roberta model. We then fine-tune the Llama 3 model using Low-Rank Adaptation (LoRA) within the Unsloth framework. Experimental results demonstrate that our enhanced RoBERTa model significantly outperforms baseline models. Furthermore, the enriched and LoRA fine-tuned Llama 3-8B model outperforms other LLMs with similar parameter scales. Our method improves MAE by 0.015 and MSE by 0.0054 on the SemEval 2018 dataset, achieving a Pearson correlation coefficient of 0.8441. On the SemEval 2017 dataset, it improves MAE by 0.0416 and MSE by 0.043, with a Pearson correlation coefficient increased to 0.8268, which demonstrates the superior predictive power and robustness of our approach.

INDEX TERMS LoRA, sentiment analysis, Llama 3, RoBERTa, adaptive weighted Huber loss function.

I. INTRODUCTION

Emotional states significantly influence human behavior and communication, serving as key indicators of our feelings and reactions [1]. With advancements in natural language processing (NLP) technologies such as emotion detection and sentiment analysis, analyzing emotions within dialogues has become feasible [2]. The core objective of sentiment analysis is to identify and extract subjective information from textual data, understanding the sentiment within the text and refining it. Traditional methods mainly focused on sentiment classification and emotion detection, which involves categorizing text into predefined sentiment categories such as positive, negative, or neutral, or classifying the emotions in the text into categories like happy, angry, or sad [3], [4], [5], [6], [7]. This technology has found widespread application in fields such as social media monitoring, customer feedback analysis, and market research [8], [9], and has been used in tasks related to mental health analysis [10], misinformation monitoring [11], and empathetic dialogue systems [12].

However, there is a paucity of research dedicated to the sentiment intensity regression task, which aims to quantify the degree of sentiment expressed in the text [13], [14]. Unlike simple sentiment classification, sentiment intensity regression requires accurately capturing subtle variations in emotional expression and handling noisy data that might obscure true emotional signals. Accurately assessing the intensity of sentiment in text, such as distinguishing between "I am very happy, today is great" and "I am happy," presents a greater challenge than discrete classification, as it provides a finer granularity of emotional features and expressions (FIGURE 1).

Pre-trained language models (PLMs) such as BERT [15] and RoBERTa [16] have demonstrated outstanding performance across numerous NLP tasks. Many studies have applied these models to sentiment analysis or emotion

The associate editor coordinating the review of this manuscript and approving it for publication was Jie Tang.

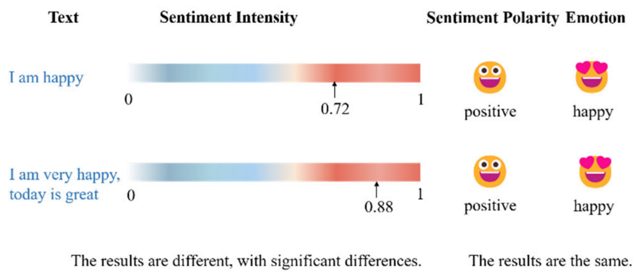

FIGURE 1. Comparison of Sentiment Intensity Regression with Other Sentiment Analysis Tasks. Sentiment intensity regression significantly highlights the differences between two example sentences. In contrast, sentiment classification and emotion detection tasks both yield the same results, "positive" and "happy," respectively, without distinguishing between the varying intensities.

detection tasks [17], [18], [19], [20], [21]. However, PLMs are limited by their model parameters and the size and quality of their training corpora, which restricts their comprehensive understanding and generalization capabilities for complex tasks, particularly in sentiment intensity regression [22], [23]. In contrast, large language models (LLMs) exhibit remarkable generalization capabilities due to their extensive parameter scales and the vastness of their training corpora. This has resulted in excellent performance in handling complex tasks [24]. Many researchers have begun exploring the application of LLMs in the field of sentiment analysis [22], [24], [25], [26], [27]. By fine-tuning open-source LLMs on sentiment analysis tasks, these models have achieved superior performance. Nonetheless, this body of research remains primarily focused on sentiment classification tasks, with relatively little attention given to sentiment intensity regression tasks [14].

To address the aforementioned challenges, this study proposes an innovative approach combining an enhanced RoBERTa model with the Llama 3 architecture to improve sentiment intensity regression tasks. Specifically, we introduce an adaptive Huber weighted loss function to the RoBERTa model, which dynamically adjusts to accommodate the magnitude of prediction errors, thereby enhancing the model's capability to handle outliers and varying error scales. Additionally, we incorporate an efficient additive attention mechanism aimed at optimizing the model's focus on critical information within the text, thereby improving the accuracy of sentiment intensity predictions.

Building on this, we utilize prompt engineering techniques on the SemEval 2017 Task 4 and SemEval 2018 Task 1 dataset, employing synonym replacement methods based on SenticNet and WordNet and An Easier Data Augmentation (AEDA) to augment text data. Additionally, we enrich the fine-tuning dataset with outputs from the enhanced RoBERTa model, ensuring the inclusion of detailed emotional expressions captured by the improved RoBERTa. This results in a high-quality sentiment intensity regression fine-tuning dataset, which is then used to fine-tune the Llama 3 model using the Unsloth framework and Low-Rank Adaptation (LoRA) techniques, thereby enhancing its performance in sentiment intensity regression tasks.

Our main contributions are as follows:

- We propose a framework based on the enhanced RoBERTa output and LoRA fine-tuning to improve the Llama 3 large language model, addressing the inaccuracy issues in emotion intensity prediction prevalent in many existing methods.
- We introduce two key enhancements to the RoBERTa model: an efficient additive attention mechanism and an adaptive weighted Huber loss function, significantly improving its performance. After applying prompt engineering to the output of the RoBERTa model, we fine-tune the Llama 3-8B model using LoRA based on this fine-tuned dataset.
- Our experimental results demonstrate significant improvements, surpassing other foundational models. Compared to the second-best model, on the SemEval 2018 dataset (with a data range of 0 to 1), our approach improves MAE by 0.015, MSE by 0.0054, and increases the Pearson correlation coefficient to 0.8441. On the SemEval 2017 dataset (with a data range of -2 to 2), MAE is improved by 0.0416, MSE by 0.043, and the Pearson correlation coefficient is improved to 0.8268.

The rest of this paper is organized as follows: Section I explains the difference between sentiment classification and sentiment intensity regression, emphasizing the neglect of the latter and the limitations of existing pre-trained language models (PLMs) and large language models (LLMs) in this area. Section II provides a comprehensive literature review on sentiment analysis and fine-tuning techniques. Section III details the methodology, including enhancing RoBERTa with efficient additive attention and adaptive weighted Huber loss, and employing data augmentation and LoRA to fine-tune the Llama 3-8B model. Sections IV and V present the experimental setup and results, discuss the findings, and propose future research directions. Section VI summarizes contributions and practical implications.

II. LITERATURE REVIEW

A. SENTIMENT ANALYSIS

In the field of sentiment analysis, existing studies have shown that linguistic knowledge, such as part-of-speech tagging and word-level sentiment polarity, is important for understanding the emotions in long texts [28], [29], [30]. Methods based on sentiment lexicons and syntactic structures have been shown to perform well in sentiment analysis.

However, although tools based on sentiment lexicons and word-level sentiment polarity, such as VADER and TextBlob, offer convenience, their effectiveness in accurately capturing emotions still needs improvement [31], [32], [33]. To address this issue, researchers have begun utilizing deep learning models and pre-trained models, training them on specific sentiment analysis datasets to enhance their sentiment analysis capabilities [20], [21], [34], [35]. Xu et al. [20] proposed BERT-PT, which improves aspect-level sentiment analysis by further training on corpora from the same domain as the downstream task. Following this, Xu et al. [21] designed DomBERT, which incorporates training samples from related domains during the pre-training phase to enhance the performance of aspect-level sentiment analysis in the target domain. Ke et al. [36] introduced word-level linguistic knowledge into pre-trained models and proposed the SentiLARE model, which employs a context-aware sentiment attention mechanism to acquire the sentiment polarity of each word along with its part-of-speech tag, achieving excellent performance across various tasks.

Liao et al. [18] proposed a multi-task model based on RoBERTa for aspect category sentiment analysis. Yin et al. [19] introduced the SentiBERT model, which focuses on the sentiment analysis domain and integrates recursive compositional structures based on BERT to better capture compositional sentiment semantics. Bello et al. [17] combined BERT with other deep learning models, such as CNN, RNN, and LSTM, to improve the capability of analyzing sentiment in short and simple texts.

B. SENTIMENT INTENSITY REGRESSION

The task of sentiment intensity regression surpasses traditional classifications of sentiment polarity or emotion categories by focusing on the nuanced measurement of sentiment intensity. As early as 2013 and 2014, tools like TextBlob and VADER were introduced, laying the foundation for sentiment analysis. VADER [32] is a rule-based tool specifically designed for analyzing sentiments in social media texts, outperforming individual human raters in assessing tweet sentiments. The 2018 SemEval Task 1 provided a detailed definition of the sentiment intensity regression task and a benchmark dataset for evaluating models in this field [37].

Xie et al. [13] proposed an attention-based CNN model for predicting sentiment intensity. Akhtar et al. [14] utilized a stacked ensemble method, combining multiple deep learning models and using a multi-layer perceptron network to predict the intensity of emotions and sentiments, achieving remarkable results in both general and financial sentiment analysis domains. Qureshi et al. [38] improved depression level estimation by concurrently learning sentiment intensity, emphasizing the importance of accurately predicting sentiment intensity for mental health assessments. Govindasamy and Palanichamy [33] used TextBlob combined with machine learning techniques to predict sentiment intensity from Twitter data, aiding in depression detection. Sergio Consoli et al. [39] highlighted the importance of fine-grained sentiment analysis in economic and financial domains. Liu et al. [27] proposed the EmoLLMs model using the AAID multi-task emotion analysis instruction dataset, demonstrating excellent performance in sentiment intensity regression tasks.

C. LARGE LANGUAGE MODELS AND FINE-TUNING

The development of Large Language Models (LLMs) has significantly revolutionized the field of Natural Language Processing (NLP). While closed-source LLMs such as GPT-4 exhibit exceptional performance across multiple domains, open-source LLMs, including the Llama series [40], [41], ChatGLM series [42], [43], Baichuan series [44], OPT series [45] and Gemma series [46], have provided researchers and developers with powerful tools and flexibility, fostering scientific research in the domain of large models. These open-source models not only demonstrate outstanding foundational performance but also offer vast potential for research based on fine-tuning techniques. Through fine-tuning, large language models are able to understand human intentions with greater precision and achieve higher performance on specific tasks.

As LLMs like Llama 2 and Llama 3 continue to evolve, their increasing parameter count escalates the energy consumption required for training. Consequently, fine-tuning techniques, which optimize these large models for specific tasks, become crucial. Houlsby et al. [47] proposed the concept of freeze-tuning, a method that preserves a model's overarching capabilities while enabling adaptation of specific layers to new tasks. Innovating within this domain, Hu et al. [48] presented model's integrity by freezing all its weights and introducing trainable low-rank matrices, providing a parameter-efficient alternative to adapt large language models effectively [3]. This method is refined through the work of Mangrulkar et al. [49], where the PEFT library offers the inclusion of adaptable adapters such as LoRA, enhancing model customizability without extensive retraining. Additionally, Dettmers et al. [50] built upon this approach with QLoRA, applying quantization to further minimize memory usage, underscoring the continual stride towards more efficient and resource-effective fine-tuning methods. Diao et al. [51] introduced LMFlow, an extensible and effective framework that facilitates full-tuning and adapter tuning for decoder-only models, enhancing model performance across various tasks.

An application area where fine-tuned LLMs are making significant inroads is sentiment analysis. With the deep contextual understanding afforded through fine-tuning, LLMs are equipped to unravel the complexities in tone, intent, and emotional expression contained within extensive textual datasets. Real-world applications of these analyses have already been seen in sectors from customer feedback analysis to broad public sentiment monitoring. Zhang et al. [22] advanced the field with their retrieval-augmented LLM for finance. Zhang et al. [10] investigated the fusion of emotional data for detecting mental illness on social networks, while Zhang et al. [24] achieved groundbreaking results with DialogueLLM in emotion recognition tasks. Complementing these, Zheng et al. [52] introduced ExTES-Llama, optimizing LLMs for emotional support in conversational AI, addressing challenges like data scarcity and training paradigms. Liu et al. [27] purposed EmoLLMs, which is the first series of open-sourced instruction-following LLMs series for comprehensive multitask affective analysis. Notably, the EmoLlama models in this series have demonstrated performance in sentiment intensity regression tasks that surpasses GPT-4.

We would like to point out that the above works all primarily focus on sentiment polarity classification, aspect-level sentiment analysis, or emotion category detection. As will be clear soon, our work differs from the existing approaches in several ways: 1) We specifically target sentiment intensity regression, capturing fine-grained emotional nuances often overlooked. 2) Our enhancements to the RoBERTa model, including the efficient additive attention mechanism and adaptive weighted Huber loss function, are designed to improve sentiment intensity regression, distinguishing our approach from models focused on sentiment polarity or emotion detection. 3) We employ prompt engineering and LoRA fine-tuning within the Unsloth framework, leveraging enriched outputs from the enhanced RoBERTa model to fine-tune Llama 3 for superior performance in complex scenarios.

III. METHODOLOGY

As illustrated in FIGURE 2, our objective is to enhance the Llama 3 model's capability for sentiment intensity regression by leveraging the improved RoBERTa method and fine-tuning through LoRA. To achieve this goal, we have delineated our approach into two distinct phases: Enhancements to the RoBERTa Model and Enhancements to the Llama 3 Model with LoRA. In phase 1, the input text undergoes processing through the enhanced RoBERTa model, which incorporates an efficient additive attention layer and an adaptive weighted Huber loss function. These advanced modifications are designed to improve the accuracy of sentiment intensity predictions by more effectively capturing the nuanced expressions within the text. In phase 2, The outputs from the enhanced RoBERTa model are then combined with augmented input data for prompt engineering. These engineered prompts and augmented texts are used to create fine-tuned datasets. Subsequently, these datasets are employed for fine-tuning the Llama 3 model using Low-Rank Adaptation (LoRA) within the Unsloth framework. This comprehensive approach ensures that the fine-tuned Llama 3 model can accurately and robustly predict sentiment intensity across diverse and complex scenarios.

A. ENHANCEMENTS TO THE RoBERTa MODEL

1) UTILIZING RoBERTa AS THE BASE MODEL

RoBERTa (Robustly Optimized BERT Pretraining Approach) is a derivative model of BERT (Bidirectional Encoder Representations from Transformers). It enhances the BERT model by adjusting the pretraining process, including modifications in training data volume, batch size, and learning rate, thereby achieving advanced capabilities in understanding contextual and semantic aspects of text. This makes RoBERTa the basic model of this study. For the task of sentiment intensity regression, RoBERTa's deep bidirectional architecture allows the model to consider both preceding and succeeding words in a sentence when deriving the contextual meaning of any given word. This bidirectional contextual understanding is more effective in assessing the sentiment intensity of sentences. Additionally, due to the scarcity of labeled data for sentiment regression tasks, RoBERTa's architecture allows for fine-tuning on relatively small task-specific datasets, which is a significant advantage for this research.

2) ADAPTIVE WEIGHT HUBER LOSS FUNCTION

In this section, we propose an adaptive weighted Huber loss function (AWHLF) to better adjust the model's performance. The adaptive weighting mechanism is designed to dynamically adjust the weight of each error based on its relative size, thus making samples with larger errors more significant and reducing the impact of samples with smaller errors.

The adaptive weighted Huber loss function builds on the adaptive weighted loss function by additionally incorporating the mechanism of the traditional Huber loss function. The Huber loss function smoothly transitions between the L2 loss for smaller errors and the L1 loss for larger errors. Making it more sensitive and robust, combined with the characteristics of adaptive weighting, the model can effectively handle both large and small error samples.

In the context of our specific sentiment regression task, where the sentiment intensity of our dataset ranges between [0, 1], this limited range can pose challenges for the traditional Huber loss function. The L2 loss component of the Huber loss is intended to make the model more sensitive to small error samples by emphasizing these small errors. However, within the [0, 1] range, the L2 loss values become even smaller, which contradicts the original design intent of emphasizing these minor deviations. To address this issue, we have modified the Huber loss function by scaling both the error and the delta parameter by a factor of n, where n represents the scaling factor. In this specific case, we have chosen 10 as the scaling factor. This adjustment ensures that small errors are appropriately emphasized, thereby enhancing the model's sensitivity and robustness across the entire range of sentiment intensities. Specifically, the computation of the adaptive weighted Huber loss function is as follows:

The first step is calculating the absolute error between predictions ($\hat{y}$) and targets (y):

$$e_i = \left| y_i - \hat{y}_i \right| \cdot n \tag{1}$$

Then the mean error ($\overline{e}$) is computed with a small constant to prevent division by zero:

$$\overline{e} = \left(\frac{1}{N} \sum_{i=1}^{N} e_i \right) + 1 \times 10^{-8} \tag{2}$$

After that, the Huber loss is calculated using the definition:

$$L_\delta\left(e_i\right) = \begin{cases} \frac{1}{2} e_i^2 & \textit{for } e_i \leq \delta, \\ \delta\left(e_i - \frac{1}{2}\delta\right) & \textit{otherwise.} \end{cases} \tag{3}$$

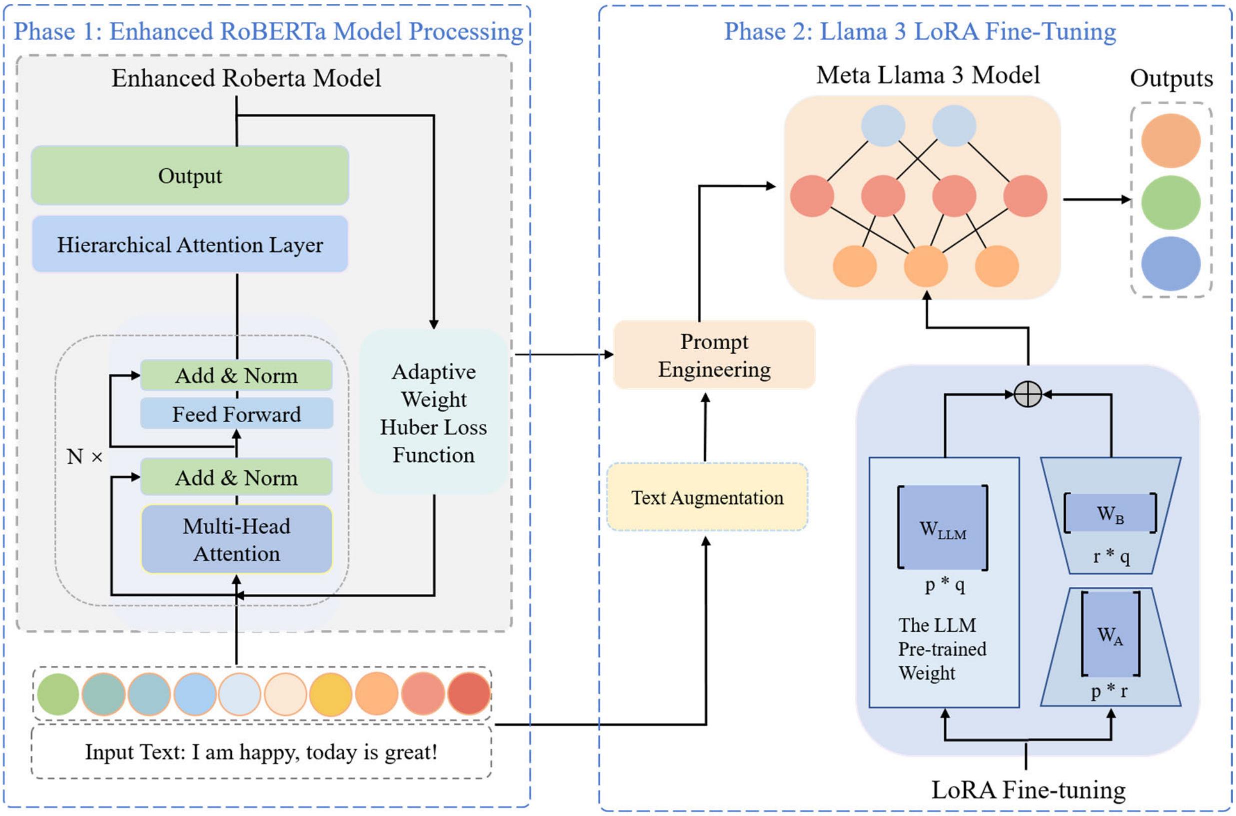

FIGURE 2. Workflow of the Enhanced Sentiment Intensity Regression Model: Input text is processed by the enhanced Roberta model with efficient additive attention and adaptive weighted Huber loss. The output is combined with augmented input data for prompt engineering, followed by fine-tuning of the Llama 3 model using LoRA within the Unsloth framework.

The weights are determined based on the relative size of each error:

$$w_i = \frac{e_i}{\overline{e}}, w_i = \frac{w_i}{\sum_{i=1}^{N} w_i} \tag{4}$$

The final step involves applying these weights to the Huber loss:

$$L_{adaptive\ weight} = \frac{1}{N}\sum_{i=1}^{N} w_i L_\delta\left(e_i\right) \tag{5}$$

Putting it all together, we get the overall loss function formula:

$$\begin{aligned} L_{adaptive\ weight} &= \frac{1}{N}\sum_{i=1}^{N} \frac{e_i/\overline{e}}{\sum_{j=1}^{N} e_j/\overline{e}} L_\delta\left(e_i\right) \\ &= \frac{1}{N}\sum_{i=1}^{N} \frac{e_i/\overline{e}}{N} L_\delta\left(e_i\right) \\ &= \frac{1}{N^2}\sum_{i=1}^{N} \frac{e_i}{\overline{e}} L_\delta\left(e_i\right) \end{aligned} \tag{6}$$

Let e_i denote the scaled absolute error for the i-th sample, where y_i and $\hat{y}_i$ represent the true and predicted sentiment intensity values, respectively. The mean of these errors is $\overline{e}$, calculated over the total number of samples, N. The Huber loss for each sample, $L_\delta(e_i)$, uses a threshold parameter δ to balance between mean absolute error and mean squared error. n represents the scaling factor. Each sample's loss is weighted by w_i, and the final adaptive Huber loss for the batch is $L_{\text{adaptive weight}}$. The detailed process flow is presented in Algorithm 1.

3) EFFICIENT ADDITIVE ATTENTION MECHANISM IN RoBERTa

In this study, we enhance the RoBERTa model's capabilities for sentiment intensity regression by integrating an efficient additive attention mechanism. This mechanism refines the RoBERTa model's output by dynamically focusing on the most relevant tokens in the text, thereby improving the accuracy of sentiment intensity predictions. The Efficient Additive Attention Mechanism operates on the sequence output from the RoBERTa model, represented as $O \in \mathbb{R}^{T\times H}$, where T is the sequence length and H is the hidden size. A learnable parameter $W \in \mathbb{R}^{H\times 1}$ is introduced to compute the attention scores $A \in \mathbb{R}^{T}$ through the equation $A = \tanh(O) \cdot W$. Employing the hyperbolic tangent (tanh) function ensures the stability of gradients during training.

To handle sequences of varying lengths, an attention mask $M \in \mathbb{R}^{T}$ is applied to zero out the scores for padded tokens.

Algorithm 1 Adaptive Weight Huber Loss Function

Input: Predicted value $\hat{y}$, Target value y, Threshold value δ, An amplification factor of type float n
Output: Weighted Loss L
var e, μ, L_{small}, L_{large}, base_loss, w, loss with weights L; array;
var i: integer;
var n: float;
for $i = 1$ to length $(\hat{y})$ **do**
  $e[i] \leftarrow |\hat{y}[i] - y[i]| \cdot n$
end for
print("Error; ", e);
$\mu \leftarrow$ mean(e)+1×10^{-8}
print("Mean Error: ", μ);
for $i = 1$ to length(e) **do**
  if $e[i] < \delta$ **then**
    $L_{\text{small}}[i] \leftarrow 0.5 \times (e[i])^2$
    base_loss[i] $\leftarrow L_{\text{small}}[i]$
  else
    $L_{\text{small}}[i] \leftarrow \delta \times (e[i] - 0.5 \times \delta)$
    base_loss[i] $\leftarrow L_{\text{large}}[i]$
  end if
end for
for $i = 1$ to length(e) **do**
  $\omega[i] \leftarrow e[i]/\mu$
end for
$w \leftarrow w/$ sum(w)
for $i = 1$ to length(e) **do**
  $L[i] \leftarrow w[i] \times$ base_loss[i]
end for
return mean(L)

The masked attention scores A' are computed as follows:

$$A'_i = \begin{cases} A_i & \text{if } M_i = 1 \\ -\infty & \text{if } M_i = 0 \end{cases} \tag{7}$$

The Softmax function then normalizes the adjusted scores to obtain the attention weights $\alpha \in \mathbb{R}^T$:

$$\alpha_i = \frac{exp\left(A'_i\right)}{\sum_{j=1}^{T} exp\left(A'_j\right)} \tag{8}$$

The final step involves computing a weighted sum of the token representations to produce a contextually enriched representation $O' \in \mathbb{R}^H$:

$$O' = \sum_{i=1}^{T} \alpha_i \cdot O_i \tag{9}$$

where the weighted output captures the essential elements of the input sequence, allowing the regression layer to make more accurate predictions of sentiment intensity.

In practical terms, this enhanced model processes input sequences by first passing them through the standard RoBERTa architecture to obtain token-level embeddings. The Efficient Additive Attention Mechanism then refines these embeddings by assigning dynamic weights to each token, reflecting their relative importance in determining sentiment intensity. The final, weighted representation is used as input to a regression layer, which predicts the sentiment intensity score. Through this approach, efficient extraction of emotional information from the text is achieved, enabling precise modeling of sentiment intensity.

B. ENHANCEMENTS TO THE LLAMA 3 MODEL AND LoRA

1) OVERVIEW OF LLAMA 3 MODEL AND ITS ARCHITECTURE

The release of the Llama 3 model marks a significant advancement in the field of open-source large language models (LLMs). Llama 3 employs a relatively standard Transformer architecture, with significant improvements over Llama 2. The model utilizes a tokenizer with a vocabulary of 128K, which greatly enhances language encoding efficiency and overall model performance. Meta optimized Llama 3's inference efficiency by employing an efficient tokenizer and grouped query attention (GQA). Notably, the Llama 3 8B model outperforms Google's Gemma 7B and Mistral 7B Instruct across several benchmarks, including MMLU, GPQA, HumanEval, and GSM-8K, which is why the Llama 3-8B model was chosen for this study.

2) TEXT DATA AUGMENTATION

Given the scarcity of annotated datasets for sentiment intensity regression tasks, this study employs synonym replacement based on SenticNet and WordNet and An Easier Data Augmentation (AEDA) technique for text data augmentation. This approach aims to expand the training dataset and enhance the model's ability to perceive sentiment regression in text, facilitating the development of more robust and accurate models.

a: SYNONYM REPLACEMENT BASED ON SENTICNET AND WORDNET

Synonym replacement is a data augmentation technique that involves replacing words in the text with their synonyms. This method increases the diversity of training examples without significantly altering the original meaning or sentiment intensity of the text. SenticNet, a sentiment analysis corpus, provides concepts and their associated semantics [53]. For selected words in the text, SenticNet identifies synonyms with similar emotional states, ensuring minimal changes in the sentiment intensity of the paragraph and preserving the original sentiment intensity as much as possible. When SenticNet does not provide suitable synonyms, WordNet is used for synonym replacement. WordNet is a lexical database that groups English words into sets of synonyms known as synsets. WordNet ensures the contextual appropriateness of synonyms, largely preserving the original sentence meaning.

$$Synonyms\,(w) = \begin{cases} SenticNet\,(w) & if\;\; w\ in\ SenticNet \\ WordNet\,(w) & otherwise \end{cases} \tag{10}$$

When performing synonym replacement, we first check the sentence for words that are present in SenticNet, excluding stop words. If such words are found, they are replaced with their SenticNet synonyms and the sentence is output. If no SenticNet words are found, we then check for synonyms of the words in the sentence using WordNet and perform the replacement accordingly.

For example:

- Original sentence: "The product is fantastic and works perfectly."
- Augmented sentence with SenticNet: "The product is wonderful and works perfectly."
- If SenticNet does not provide suitable synonyms, using WordNet: "The product is good and works perfectly."

b: AN EASIER DATA AUGMENTATION (AEDA)

AEDA enhances training data by randomly inserting punctuation marks into the text, offering the dual advantages of quick implementation and preservation of the original word order [54]. This method ensures the syntactic and semantic integrity of sentences, minimizing the risk of altering the original meaning and sentiment, which is a common drawback in other data augmentation techniques in the sentiment intensity regression task.

The process of AEDA involves the following steps: First, sentences longer than one word are selected for augmentation. For each selected sentence, two punctuation marks are randomly inserted within the sequence, ensuring sufficient augmentation while avoiding excessive noise that could negatively impact the model. The insertion points are chosen at random to maintain the natural flow of the text. The punctuation marks are selected from a predefined set, including "." (period), "," (comma), ";" (semicolon), ":" (colon), "-" (hyphen), and "..." (ellipsis), as their impact on sentiment intensity is relatively minimal. Punctuation marks such as "!" (exclamation mark) and "?" (question mark) are excluded due to their potential to significantly affect the sentiment intensity, given the high sensitivity of sentiment intensity regression tasks to textual sentiment.

For example, consider the original sentence: "The product is fantastic and works perfectly." After applying AEDA, the augmented sentence might be: "The product is fantastic, and works - perfectly."

By employing this method, we increase the number of training samples, enhancing the robustness and generalization capabilities of the model. This approach ensures that the original sequence information and emotional state are preserved to the greatest extent possible, thereby improving the model's ability to perform sentiment intensity regression tasks effectively.

3) LLAMA 3 LORA FINE-TUNING BASED ON UNSLOTH

The resource and computational power required for pre-training large language models like Llama 3 on specific tasks are immense, making it impractical to fully fine-tune all model parameters. To improve efficiency and reduce resource consumption, we adopt the Unsloth framework to optimize the Low-Rank Adaptation (LoRA) method for fine-tuning the Llama 3 model on sentiment regression tasks.

LoRA (Low-Rank Adaptation) is a technique that fine-tunes pre-trained language models by introducing low-rank matrices into the weight matrices of the model. LoRA fine-tuning is achieved by inserting low-rank matrices into certain layers of the model to efficiently adjust parameters, thereby reducing computational and memory overhead while maintaining the model's expressiveness. The basic principle is as follows:

Let $W_0 \in \mathbb{R}^{d\times k}$ be the pre-trained weight matrix, and the weight update ΔW is expressed as:

$$W = W_0 + \Delta W = W_0 + BA \tag{11}$$

where $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times k}$, with $r \ll \min(d, k)$. During training, W_0 remains fixed and does not receive gradient updates, while A and B are trainable parameters.

For input x, the original forward pass is:

$$h = W_0 x \tag{12}$$

Besides, the following modifications are made in the forward pass to accommodate downstream tasks:

$$h = W_0 x + \Delta W x = W_0 x + BAx \tag{13}$$

Due to the smaller LoRA weight matrices, the performance is greatly improved by correctly placing parentheses when combining with the multi-level matrix multiplication of Llama 3 weight matrices. The specific technical principles can be shown in detail in FIGURE 3.

For this experiment, we utilized the Unsloth framework, which is a LoRA fine-tuning framework based on OpenAI's Triton language. This framework significantly accelerates the training of large language models (LLMs) by rewriting kernels. The open-source version can double the fine-tuning speed while reducing memory usage by 50%. Unsloth optimizes the LoRA training process through manual matrix differentiation. Given the smaller LoRA weight matrices, performance is greatly improved by correctly placing parentheses when combining with the multi-level matrix multiplication of LLaMA 3 weight matrices. Additionally, Unsloth introduces Triton to implement manual automatic differentiation and chain matrix multiplication optimization, reducing floating-point operations and speeding up LoRA training.

4) FROM ROBERTA TO LLAMA 3 CROSS-MODEL ENHANCEMENT

The refined output from the improved RoBERTa model serves as crucial input to enhance the Llama 3 model through a cross-model enhancement method. This integration leverages the detailed predicted results generated by RoBERTa to enrich Llama 3's training data, thereby improving its performance in sentiment intensity regression tasks. The improved

TABLE 1. Prompt of LLama 3 model.

Description	Details On SemEval 2018 Dataset
Task	Evaluate the emotional valence of the following text and assign a single score from 0 (most negative) to 1 (most positive), Note that the emotional valence cannot be 0 or 1.
Text	{tweet}
Enhanced RoBERTa Model Result	{RoBERTa _output}
Note	Please note that not all sentences have an emotional valence of 0 or 1. Please objectively evaluate the emotional valence of the sentence and provide a floating-point number between 0 and 1 to better measure the emotional valence of the sentence, such as 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, etc.
Intensity Score	{score}

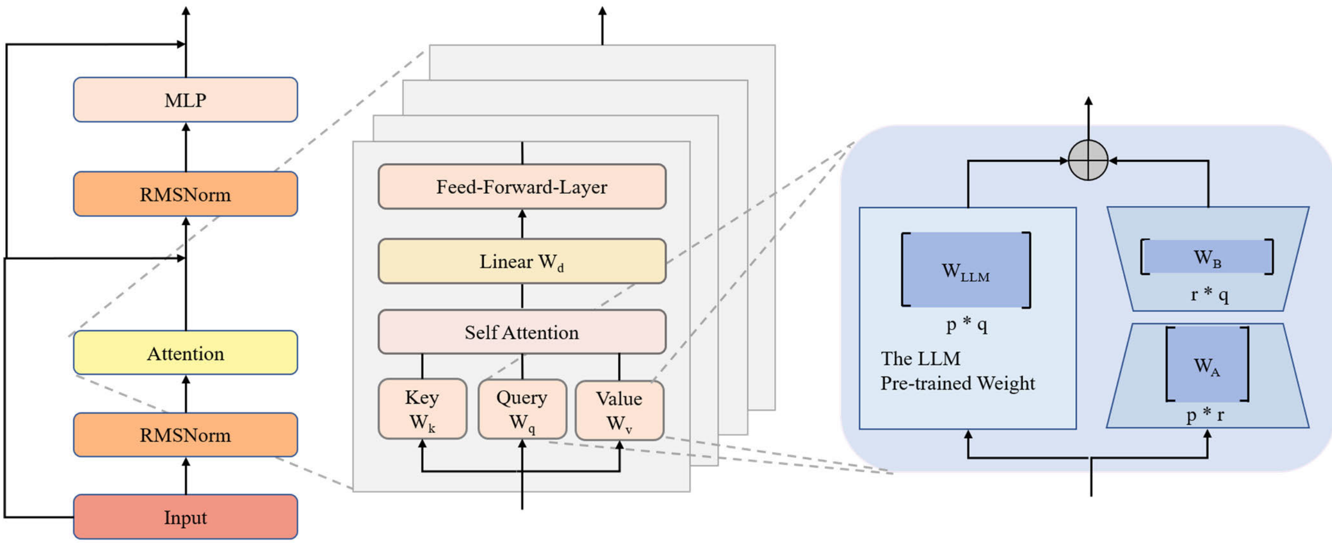

FIGURE 3. Low-Rank Adaptation (LoRA) Method in Llama 3's Transformer Module. This figure demonstrates the application of the Low-Rank Adaptation (LoRA) method in the attention layers of the Llama 3 model's transformer module. The left side depicts the transformer architecture, highlighting the input, RMS normalization, attention mechanism, feed-forward layer, and MLP. LoRA is specifically applied to the Query (Q) and Value (V) projection matrices in the attention layer. The right side details the LoRA process, where the original pre-trained weight matrix W_{llm} (with dimensions $p \times q$) is decomposed into two smaller matrices, W_A (dimensions $p \times r$) and W_B (dimensions $r \times q$). During fine-tuning, these smaller matrices are updated instead of the large matrix W_{llm}, significantly reducing computational complexity. The adapted weight matrix is then reconstructed as $W = W_{llm} + W_A W_B$. This method allows efficient and scalable fine-tuning of the large language model.

RoBERTa model initially processes tweets or textual data, assigning each a sentiment score that reflects the intensity of emotion conveyed. First, the pre-trained embeddings from RoBERTa are used to convert the tokens of the tokenized text into embedding vectors. Then, these embedding vectors are processed through RoBERTa's transformer layers, which include multi-head attention mechanisms and feed-forward neural network layers to understand the context. These transformer layers are stacked multiple times (N times), with each layer including addition and normalization steps. Next, the model further processes the features through an efficient additive attention mechanism. Finally, the sentiment scoring layer applies a linear transformation and Softmax function to the final hidden states from RoBERTa to output the sentiment intensity. During the training iterations, an adaptive weighted Huber loss function is employed to optimize the model parameters, ensuring the accuracy of the sentiment intensity. These scores are not treated as final but are used as informed suggestions within the Llama 3 training process.

RoBERTa's contextual embeddings are used to enhance Llama 3's input prompts. By integrating these embeddings, Llama 3 can leverage the detailed semantic and sentiment information captured by RoBERTa, resulting in more accurate sentiment intensity predictions. The prompt used for Llama 3 is designed as shown in TABLE 1.

$$Y_{LLama3} = LLama3\left(P\left(E_{RoBERTa}\right)\right) \tag{14}$$

where Y_{LLama3} represents the output of Llama 3, and $P(E_{RoBERTa})$ denotes the enriched prompts created from RoBERTa outputs.

The enhanced Llama 3 model is fine-tuned using LoRA (Low-Rank Adaptation), which adapts the model to the specific task of sentiment intensity regression with reduced computational cost. This fine-tuning process involves updating a subset of the model's parameters based on the enriched training data.

$$\theta^{*}_{LLama3} = FineTune\left(\theta_{LLama3}, P\left(E_{RoBERTa}\right)\right) \tag{15}$$

where θ^*_{LLama3} represents the fine-tuned parameters of the Llama 3 model.

By combining the advanced predicted results from the improved RoBERTa model with the robust architecture of Llama 3, this cross-model enhancement method significantly improves the accuracy and robustness of sentiment intensity regression. This innovative approach leverages the strengths of both models, resulting in superior performance in predicting nuanced emotional intensities.

IV. EXPERIMENT AND RESULT

A. DATASET

In this study, we utilize two publicly available sentiment intensity regression datasets: the SemEval 2017 Task 4 Dataset [55] and the SemEval 2018 Task 1 Dataset [37]. Table 2 presents the fundamental statistics for these datasets. Below, we provide a brief introduction to each dataset.

TABLE 2. Statistics of SemEval 2018 Task 1 and SemEval 2017 Task 4 Datasets.

Dataset	SemEval 2018 Task 1			SemEval 2017 Task 4		
	Size	Mean Intensity	Standard Deviation	Size	Mean Intensity	Standard Deviation
Train	1181	0.5005	0.2093	3500	0.3580	0.7618
Dev	449	0.4846	0.2259	1000	0.3110	0.7555
Test	937	0.5202	0.2173	500	0.1660	1.0201

1) SemEval-2017 TASK 4 DATASET

The SemEval-2017 Task 4 dataset is a popular dataset for sentiment analysis, particularly focusing on sentiment classification and regression tasks in Twitter data. The dataset is divided into 5 subtasks, including Subtask CE, which involves rating the sentiment of text on a scale from -2 to 2, where -2 indicates very negative sentiment and 2 indicates very positive sentiment. In this study, we utilize the English subset of the dataset, specifically selecting 5,000 samples. These samples are split into training, development, and test sets in a 7:2:1 ratio to ensure a robust evaluation of sentiment analysis models.

2) SemEval-2018 TASK 1 DATASET

The SemEval-2018 Task 1 dataset is designed to measure the intensity of emotions and sentiments expressed in tweets, offering a detailed understanding of affective states in social media texts. This dataset is manually annotated to ensure high-quality and reliable data, providing a robust foundation for evaluating the performance of sentiment analysis models. Our focus is on the Valence or Sentiment Intensity Regression task (V-reg), where each tweet is annotated with real-valued scores representing the intensity of emotions and sentiments, ranging from 0 (least intense) to 1 (most intense). This study focuses on the English subset of the SemEval-2018 Dataset Task 1.

B. BASELINES

To validate the effectiveness of the integrated RoBERTa-Llama 3 model, various baselines and comparison models were utilized.

1) TEXTBLOB

A simple library for processing textual data, TextBlob provides a straightforward API for NLP task. For sentiment analysis, it assigns a polarity score ranging from -1 to 1, which we have adjusted to 0 to 1 for our study.

2) VADER (VALENCE AWARE DICTIONARY AND SENTIMENT REASONER)

A lexicon and rule-based tool designed for social media sentiments, VADER provides a compound sentiment score from -1 (most negative) to $+1$ (most positive). We have adjusted this score to 0 to 1 for consistency.

3) RoBERTa(PRE-TRAINED)

An optimized variant of BERT, RoBERTa improves with more training data, larger batch sizes, and longer training times, achieving state-of-the-art performance on various NLP benchmarks. For this study, we choose the model pretrained on a large corpus of social media text.

4) XLNet

A generalized autoregressive pretraining method, XLNet leverages both autoregressive and autoencoding models with a permutation-based training objective. It captures bidirectional context and achieves competitive performance in sentiment analysis.

5) ALBERT

A lighter and faster version of BERT, ALBERT (A Lite BERT) uses parameter reduction techniques to lower memory consumption and increase training speed while maintaining performance. It achieves competitive results in various NLP tasks, including sentiment analysis.

6) GEMMA-7B

The Gemma-7B model is part of an open model series based on Google's Gemini models. Gemma demonstrates enhanced performance in various tasks, including conversational AI, mathematical problem-solving, and code generation. It achieves high scores on benchmarks such as MMLU (64.3%) and MBPP (44.4%). Gemma-7B outperforms models like Llama-2-7B and Mistral-7B in multiple tasks, making it a robust and versatile tool for a wide range of NLP applications.

7) EmoLlama-7B

EmoLLM, trained on the first multi-task affective analysis instruction dataset, outperforms all other open-source LLMs and sentiment analysis tools in various affective analysis tasks, even surpassing some capabilities of ChatGPT and

GPT-4. The sub-model EmoLlama-7B balances model size and performance, ensuring efficient operation while maintaining strong performance across different tasks.

8) Llama 3-8B

Llama 3 is the most advanced open-source large language model, and its sub-model, Llama 3-8B, demonstrates exceptional performance at the same parameter level. It excels in various tasks, showcasing outstanding capabilities.

C. EXPERIMENTAL SETTING

In the field of sentiment intensity regression, there is currently no universally accepted standard metric. Therefore, based on common practices in other sentiment intensity regression tasks and standard metrics used in regression tasks, this study employs Mean Squared Error (MSE), Mean Absolute Error (MAE), and Pearson Correlation Coefficient as evaluation metrics. By using these metrics, we provide a comprehensive evaluation of model performance in sentiment intensity regression tasks.

We trained a customized RoBERTa model with an efficient additive attention mechanism and a regression layer for predicting sentiment intensity scores. An adaptive weight Huber loss function was used, with the delta set to 1.2, and the model was trained for 50 epochs with a batch size of 128 for training and 64 for evaluation. The AdamW optimizer with a weight decay of 0.01 was employed, and the learning rate was dynamically adjusted. Text data was tokenized to a maximum sequence length of 512.

For the Llama 3-8B model, optimized for 4-bit precision, we used a specific prompt template to enhance emotional valence evaluation based on the RoBERTa model's output. Training parameters included a batch size of 2, gradient accumulation steps of 4, warmup steps of 5, and a maximum of 150 training steps. The learning rate was set to 8e-5 with the AdamW optimizer in 8-bit precision. Logs were recorded at every step, with periodic evaluations. We utilized LoRA technology with the Unsloth framework for efficient adaptation, setting the rank r to 16.

The training was performed on Kaggle's GPU environment equipped with Nvidia Tesla P100 GPUs, which have 16 GB of memory, and required careful management of computational resources due to their limited memory.

D. RESULTS

Table 3 shows the comparative results on SemEval 2018 Task 1 Dataset and SemEval 2017 Task 4 Dataset. Below, we will explain our experimental results from six perspectives: pre-trained models, large language models, ablation studies, model replacement studies, impact of text augmentation techniques, and visualization.

1) RESULTS OF THE PRE-TRAINED MODEL

From the perspective of pre-trained models, we can observe that the RoBERTa model, after incorporating an adaptive weighted Huber loss function and an efficient additive attention mechanism, performs exceptionally well on both tasks. In the SemEval 2018 task, the improved RoBERTa model achieved an MAE of 0.0927, an MSE of 0.0164, and a Pearson correlation coefficient of 0.8427. These results are superior to all other baseline models across all three-performance metrics. In the SemEval 2017 task, the improved RoBERTa model showed an MAE of 0.3838, an MSE of 0.3614, and a Pearson correlation coefficient of 0.8105, demonstrating excellent sentiment intensity regression capabilities. The training loss for the enhanced RoBERTa model, as illustrated in FIGURE 4, shows a rapid convergence after introducing the adaptive weighted Huber loss function and the efficient additive attention mechanism with performance improving from 30 to 50 epochs, highlighting the effectiveness of these enhancements.

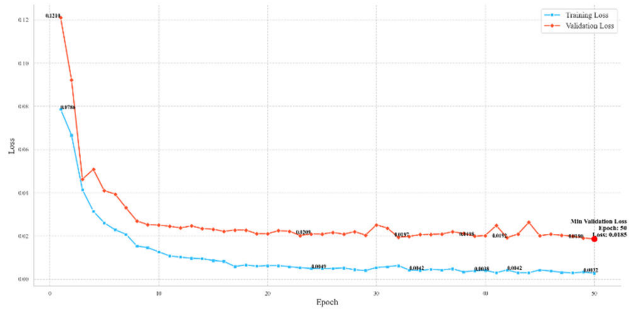

FIGURE 4. Training and validation loss curves for the enhanced Roberta model. The graph displays the training loss (blue) and validation loss (red) across 50 epochs.

The second-best baseline model in both tasks is XLNet. In the first task, XLNet achieved an MAE of 0.1067, which is 0.014 higher than the improved RoBERTa model. The MSE for XLNet was 0.0191, which is 0.0027 higher than the improved RoBERTa model. The Pearson correlation coefficient for XLNet was 0.7857, which is 0.057 lower than the improved RoBERTa model. In the second task, XLNet showed an MAE that is 0.0416 higher than our improved model.

From this series of experiments using unsupervised methods and pre-trained models, it can be observed that introducing the adaptive weighted Huber loss function and efficient additive attention significantly enhances the performance of the RoBERTa model, making it excel in sentiment intensity regression tasks.

2) RESULTS OF LARGE LANGUAGE MODEL

From the perspective of large language models, this section explores the performance of the Llama 3 large language model enhanced with RoBERTa and fine-tuned with LoRA, along with other baseline large language models, on the task of sentiment intensity regression. Observing the results in TABLE 3, we can see that the Llama 3 large language model, enhanced with RoBERTa and fine-tuned with LoRA, exhibited excellent performance on sentiment intensity regression tasks across both datasets: On the SemEval 2018 dataset,

TABLE 3. Results on SemEval 2018 task 1 dataset and SemEval 2017 task 4 datasets.

Model	SemEval 2018 Task 1			SemEval 2017 Task 4		
	MAE	MSE	Pearson Correlation (p)	MAE	MSE	Pearson Correlation (p)
VADER	0.1746	0.0473	0.7127	0.7132	0.8106	0.6281
TextBlob	0.1682	0.0449	0.5076	0.8874	0.9799	0.372
RoBERTa-pre-train	0.168	0.041	0.284	0.9234	1.0536	0.4975
ALBERT	0.1096	0.0199	0.7607	0.5165	0.4984	0.7108
XLNet	0.1067	0.0191	0.7857	0.4194	0.3691	0.8029
Att + AWHLF + RoBERTa	0.0927	0.0164	0.8427	0.3838	0.3614	0.8105
Gemma	0.2556	0.0987	0.0173	0.862	1.7344	-0.0366
Original Llama 3	0.2386	0.0874	0.0328	1.108	1.7400	0.0506
Llama 3 with LoRA Fine-tuning	0.2254	0.0781	0.0268	1.097	1.6755	0.0922
EmoLLM	0.1220	0.0272	0.704	0.649	0.6037	0.6762
Ours	**0.0917**	**0.0137**	**0.8441**	**0.3778**	**0.3261**	**0.8286-**

our model achieved an MAE of 0.0917, an MSE of 0.0137, and a Pearson correlation coefficient of 0.8441, significantly surpassing all baseline models in each performance metric. On the SemEval 2017 dataset, this method also demonstrated outstanding performance, with an MAE of 0.3778, an MSE of 0.3261, and a Pearson correlation coefficient of 0.8286.

The results indicate that the enhanced Llama 3 model, which combines RoBERTa output with efficient additive attention and an adaptive weight Huber loss function, outperforms all other models in predicting sentiment intensity. In the SemEval 2018 Task 1 Dataset, the enhanced Llama 3 model achieved the lowest MAE of 0.0917, which is 0.0303 lower than the second-best model, EmoLLM, with an MAE of 0.1220. This substantial reduction in MAE demonstrates the enhanced Llama 3 model's superior accuracy in predicting emotion intensities. In terms of MSE, the enhanced Llama 3 model exhibited the lowest value of 0.0137, which is 0.0135 lower than EmoLLM's MSE of 0.0272. This significant decrease in MSE underscores the model's robustness in minimizing prediction errors. Furthermore, the enhanced Llama 3 model achieved a Pearson correlation coefficient of 0.8441, indicating a strong positive correlation between the predicted and actual sentiment intensities. This correlation is 0.1401 higher than that of EmoLLM-Llama-7B, which has a Pearson correlation coefficient of 0.704, illustrating the enhanced model's improved predictive power and consistency with actual values. Comparatively, the original Llama 3 model showed a high MAE (0.2386) and MSE (0.0874), along with a lower Pearson correlation coefficient (0.0328), indicating poor performance in predicting emotion intensities. The Llama 3 model with LoRA fine-tuning improved over the original model but still lagged significantly behind the enhanced version, with an MAE of 0.2254, MSE of 0.0781, and a Pearson correlation coefficient of 0.0268. The Gemma model performed far from our expectation, with an MAE of 0.2556, MSE of 0.0987, and a Pearson correlation coefficient of 0.0173, indicating its inadequacy in this task. This also shows that large language models do not perform well in the specific task of emotional intensity regression without targeted fine-tuning of a large amount of data. They are not as good as pre-trained models that are fine-tuned for specific tasks.

The experimental results clearly demonstrate the effectiveness of enriching Llama 3 training data with output from the enhanced RoBERTa model and applying LoRA fine-tuning. The robust performance of this method is due, on the one hand, to Llama 3 being the most comprehensive open-source large model, which excels in tasks requiring strong text understanding and logical reasoning abilities, such as sentiment intensity regression. On the other hand, it also validates the effectiveness of enriching Llama 3 training data with output from the improved RoBERTa model and applying LoRA fine-tuning to Llama 3 for sentiment intensity regression tasks.

3) ABLATION STUDIES

In this section, we present the results of ablation studies conducted to understand the contribution of each enhancement component in the improved RoBERTa and Llama 3 models. By systematically removing key components, we assess the impact of each enhancement on the model's performance in sentiment intensity regression. The performance metrics of these models are summarized in the TABLE 4 below:

TABLE 4. Evaluation results of ablation study.

Model	MAE	MSE	Pearson Correlation
Att+AWHLF+RoBERTa+Llama 3	0.0917	0.0137	0.8441
Att + AWHLF + RoBERTa	0.0927	0.0164	0.8427
AWHLF+ RoBERTa	0.1004	0.0168	0.8407
Att + RoBERTa	0.0954	0.0145	0.8471
RoBERTa	0.1048	0.0180	0.8029

The ablation studies reveal the individual contributions of the efficient additive attention mechanism and the adaptive Huber weight loss function to the model's overall performance. The complete model, which integrates both enhancements and enriches the Llama 3 model with improved RoBERTa output, achieved the best performance with an

MAE of 0.0917, MSE of 0.0137, and a Pearson correlation coefficient of 0.8441. Using only the enhanced RoBERTa model without Llama 3 slightly increased MAE to 0.0927 and MSE to 0.0164, with a minor decrease in Pearson correlation to 0.8427, indicating Llama 3's additional refinement of predictions. Removing the efficient additive attention mechanism (Drop-Attention) increased MAE to 0.1004, MSE to 0.0168, and slightly decreased Pearson correlation to 0.8407, demonstrating its significant role in reducing prediction errors. Similarly, removing the adaptive Huber loss function (Drop-Loss) resulted in an MAE of 0.0954, MSE of 0.0145, and Pearson correlation of 0.8471, showing its critical role in minimizing errors. The base RoBERTa model's performance (MAE: 0.1048, MSE: 0.0180, Pearson: 0.8029) underscores the improvements brought by these enhancements.

The ablation studies provide clear evidence of the effectiveness of the efficient additive attention mechanism and the adaptive Huber weight loss function in enhancing the performance of the RoBERTa and Llama 3 models for sentiment intensity regression. The combined enhancements result in the best performance, with the lowest MAE and MSE values and the highest Pearson correlation coefficient. Removing either the efficient additive attention mechanism or the adaptive loss function leads to noticeable declines in performance, highlighting their critical contributions.

4) MODEL REPLACING STUDIES

To further validate the robustness and effectiveness of our proposed enhancements, we conducted a model replacement experiment by substituting the RoBERTa base model with BERT while keeping the efficient additive attention mechanism and adaptive weighted Huber loss function (AWHLF) unchanged. The method of enhancing and fine-tuning the Llama 3 model also remained the same. This experiment aims to demonstrate that our method remains superior to other baseline models even when the underlying pre-trained model is replaced.

The experiment was designed and conducted on the SemEval 2018 Task 1 Dataset, with specific results shown in the TABLE 5:

TABLE 5. Result of model replacing study.

Model	MAE	MSE	Pearson Correlation
Att+AWHLF+RoBERTa+Llama 3	0.0917	0.0137	0.8441
Att + AWHLF + RoBERTa	0.0927	0.0164	0.8427
Att + AWHLF + BERT+ Llama 3	0.0943	0.0150	0.8303
Att + AWHLF + BERT	0.0949	0.0150	0.8300

The experimental results indicate that, after replacing RoBERTa with BERT, the enhanced pre-trained model achieved an MAE of 0.0949, an MSE of 0.0150, and a Pearson correlation coefficient of 0.83 on the dataset. Although these results are slightly inferior to the RoBERTa model, they are still superior to a range of baseline models. Furthermore, when fine-tuned using the Llama 3 model, the performance is further improved, achieving an MAE of 0.0943, and a Pearson correlation coefficient of 0.8303.

This suggests the effectiveness of RoBERTa's enhancements over the BERT model to some extent. These findings suggest that the core strength lies in our enhancement methods rather than the specific choice of RoBERTa or BERT as the base model. When substituting RoBERTa with BERT, our method still achieves superior performance compared to baseline models across all evaluated metrics. This indicates the robustness and generalizability of our enhancements, making them valuable for improving sentiment intensity regression tasks in various contexts.

5) IMPACT OF TEXT AUGMENTATION TECHNIQUES

From the results presented in FIGURE 5, it is evident that in the absence of text augmentation techniques such as AEDA and synonym replacement based on SenticNet and WordNet, the benchmark results for the models were as follows: MAE was 0.0958, MSE was 0.0155, and Pearson Correlation was 0.8238. However, upon applying text augmentation techniques, there was a noticeable improvement in performance; MAE decreased to 0.0917, MSE to 0.0137, and Pearson Correlation increased to 0.8441.

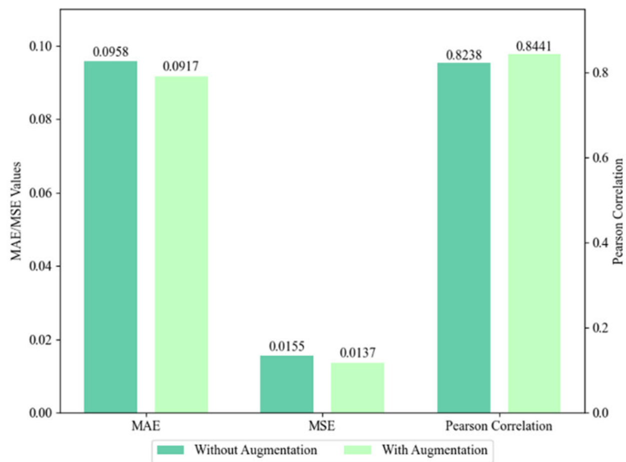

FIGURE 5. Impact of Text Augmentation on Model Performance.

These figures underscore the critical role that text augmentation techniques play in enhancing model efficiency. AEDA, by generating adversarial samples, enhances the model's robustness, while synonym replacement based on SenticNet and WordNet enriches the semantic diversity of the samples. The application of these techniques increases the number of training samples available for fine-tuning, thereby boosting the model's robustness and generalization capabilities, ultimately leading to improved overall performance.

This enhancement solidifies the model's ability to comprehend and process input data more effectively, ensuring consistently accurate predictions regardless of dataset variability.

6) VISUALIZATION

Using the mainstream text attention visualization tool, TAHV, [56], we visualized the attention scores of the RoBERTa model. The detailed results are presented in FIGURE 6.

From FIGURE 6, it can be observed that the improved RoBERTa model significantly increases attention to emotional keywords. For example, in the third sentence, "saddest" and "sadness" and in the fourth sentence, "sad," all received high attention scores. These words are crucial elements of emotional expression. The model accurately identifies these emotionally charged words and assigns them higher weights, indicating that the introduction of efficient additive attention mechanisms effectively enhances the accuracy and robustness of emotion recognition.

The visualization results in FIGURE 6 also demonstrate a close relationship between attention scores and the overall emotional intensity of the text. The model assigns higher attention scores to words with strong emotional expressions, which aligns well with the overall emotional intensity of the sentences. For instance, in the second sentence, "happy" and "awesome" not only received high attention scores but also reflected the positive emotional intensity of the entire sentence. Through this approach, the model effectively maps the attention scores of emotional keywords to the overall emotional intensity, allowing for a more nuanced understanding and processing of complex emotional expressions.

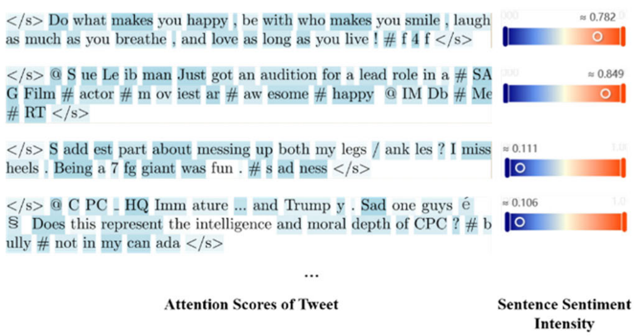

FIGURE 6. **Visualization of attention scores of the RoBERTa model.**

V. DISCUSSION

A. DISCUSSION OF RESULTS

The results of this study underscore the effectiveness of the enhanced RoBERTa and Llama 3 models in sentiment intensity regression. The improvements made to the RoBERTa model, combined with enriching the Llama 3 model's training data with output from the enhanced RoBERTa model and applying LoRA fine-tuning, demonstrate the efficacy of these methods in specific sentiment regression tasks. The findings from ablation studies further validate these results.

Regarding the performance enhancement of the RoBERTa model, the adaptive Huber weight loss function improves robustness to outliers and sensitivity to small errors, effectively reducing prediction errors. Additionally, the introduction of the attention mechanism enhances the model's alignment with human emotional understanding, thereby improving the overall performance of the model.

The effectiveness of enriching the Llama 3 model's training data with output from the enhanced RoBERTa model and applying LoRA fine-tuning can be attributed to two main factors. First, Llama 3 itself is a large language model that has undergone extensive training on vast textual datasets. This extensive training endows Llama 3 with a profound ability to understand hidden information in text and capture emotional nuances more effectively when guided, showcasing the advantage of large language models with substantial training data and parameters, which offer superior generalization and comprehension capabilities compared to pre-trained models.

Second, the improved output from the RoBERTa model significantly aids Llama 3. It is well-known that large language models can suffer from the "large model hallucination" problem. Providing improved RoBERTa outputs as guidance helps mitigate this issue by directing the large language model more precisely towards the sentiment intensity regression task. This approach reduces instances of hallucination and enhances the model's performance by allowing it to more accurately identify and interpret emotional content. When faced with uncertainty, the Llama 3 model benefits from relying on the refined cues provided by the enhanced RoBERTa, leading to improved accuracy and precision in problem-solving.

B. APPLICATION OF SENTIMENT INTENSITY REGRESSION

Compared to traditional sentiment analysis classification tasks, sentiment intensity regression provides more detailed sentiment analysis results. Traditional sentiment classification can only simply judge whether the sentiment is positive, negative, or neutral, which can be too general in decision-making and may miss some important information. In contrast, sentiment intensity regression can quantify the strength of sentiment, ranging from the most negative (0) to the most positive (1), thereby capturing more subtle emotional changes. This helps businesses to more accurately understand customer needs and market sentiments, enabling more informed and targeted decision-making.

Specifically, sentiment intensity regression has practical applications in multiple fields, including customer feedback analysis, social media monitoring, and misinformation detection.

In customer feedback analysis, sentiment intensity regression helps businesses gain deeper insights into the emotional intensity of customer comments. For comments with moderate intensity, such as "This product is okay," they can be marked for regular follow-up, while comments with extreme intensity (whether very negative or very positive) should be prioritized. Very negative comments may require immediate customer support action, while very positive comments can lead to customer rewards or recommendations.

By evaluating sentiment intensity, businesses can optimize service and product quality, thereby improving overall business performance.

In social media monitoring, sentiment intensity regression allows businesses to identify strong public reactions to events or brands. For example, when a large number of very negative comments are detected, businesses can respond quickly by issuing official statements or taking measures to control public opinion. Conversely, for comments with low but not extremely negative intensity, businesses can conduct long-term monitoring and analysis to develop long-term brand strategies.

In misinformation detection, very negative information is more likely to attract widespread attention. Sentiment intensity regression can help identify and respond to such information earlier, preventing the spread of misinformation.

C. ADVANTAGES AND LIMITATIONS

This study successfully integrates an efficient additive attention mechanism and an adaptive Huber weight loss function into the RoBERTa model, significantly improving its performance in sentiment intensity regression tasks. By enriching the Llama 3 model's training data with output from the enhanced RoBERTa model and applying LoRA fine-tuning, we demonstrate substantial improvements in predicting sentiment intensities, achieving results that surpass baseline models. The detailed ablation studies conducted provide clear evidence of the critical contributions of each enhancement component, underscoring the robustness and effectiveness of the proposed methods.

However, there are limitations to this research. Due to the limited availability of datasets specifically for sentiment intensity regression, this study only utilized two datasets for experiments. Traditional sentiment analysis datasets like IMDB, which label sentiment as binary (positive as 1, negative as 0), differ significantly from the needs of sentiment intensity regression tasks. Using these datasets for training would likely result in models that output values close to 0 and 1, conflicting with the goals of sentiment intensity regression. Additionally, due to constraints in experimental equipment, this study did not include fine-tuning experiments on large parameter models such as the Llama 3-70B model. Consequently, the results do not represent the extreme performance capabilities of large language models but rather those of smaller parameter models like the 8B and 7B models.

VI. CONCLUSION

This study demonstrates significant advancements in sentiment intensity regression by enhancing the RoBERTa model with an efficient additive attention mechanism and an adaptive Huber weight loss function. These enhancements substantially improved the model's performance, achieving greater accuracy and reliability in predicting emotion intensities. Furthermore, enriching the Llama 3 model's training data with outputs from the improved RoBERTa model and applying LoRA fine-tuning further elevated performance, surpassing baseline models. This work enhances the model's alignment with human emotional understanding, providing more accurate and fine-grained sentiment analysis for applications such as customer feedback analysis, social media monitoring, and mental health assessments.

Future research should utilize diverse datasets and larger models like Llama 3-70B to improve the generalizability and robustness of sentiment intensity regression, while testing enhanced models in real-world applications and exploring multimodal data integration to advance sentiment analysis.

REFERENCES

[1] B. Liu, *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge, U.K.: Cambridge Univ. Press, 2020.

[2] N. M. Hakak, M. Mohd, M. Kirmani, and M. Mohd, "Emotion analysis: A survey," in *Proc. Int. Conf. Comput., Commun. Electron. (Comptelix)*, Jul. 2017, pp. 397–402, doi: 10.1109/COMPTELIX.2017.8004002.

[3] A. Yadav and D. K. Vishwakarma, "Sentiment analysis using deep learning architectures: A review," *Artif. Intell. Rev.*, vol. 53, no. 6, pp. 4335–4385, Aug. 2020, doi: 10.1007/s10462-019-09794-5.

[4] M. Birjali, M. Kasri, and A. Beni-Hssane, "A comprehensive survey on sentiment analysis: Approaches, challenges and trends," *Knowl.-Based Syst.*, vol. 226, Aug. 2021, Art. no. 107134, doi: 10.1016/j.knosys.2021.107134.

[5] S. Zad, M. Heidari, J. H. J. Jones, and O. Uzuner, "Emotion detection of textual data: An interdisciplinary survey," in *Proc. IEEE World AI IoT Congress (AIIoT)*. Seattle, WA, USA: IEEE, May 2021, pp. 255–261, doi: 10.1109/AIIoT52608.2021.9454192.

[6] S. J. R. K. Padminivalli V, M. V. P. C. S. Rao, and N. S. R. Narne, "Sentiment based emotion classification in unstructured textual data using dual stage deep model," *Multimedia Tools Appl.*, vol. 83, no. 8, pp. 22875–22907, Aug. 2023, doi: 10.1007/s11042-023-16314-9.

[7] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas, "Sentiment strength detection in short informal text," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 61, no. 12, pp. 2544–2558, Dec. 2010, doi: 10.1002/asi.21416.

[8] P. K. Jain, R. Pamula, and G. Srivastava, "A systematic literature review on machine learning applications for consumer sentiment analysis using online reviews," *Comput. Sci. Rev.*, vol. 41, Aug. 2021, Art. no. 100413, doi: 10.1016/j.cosrev.2021.100413.

[9] P. Sudhir and V. D. Suresh, "Comparative study of various approaches, applications and classifiers for sentiment analysis," *Global Transitions Proc.*, vol. 2, no. 2, pp. 205–211, Nov. 2021, doi: 10.1016/j.gltp.2021.08.004.

[10] T. Zhang, K. Yang, S. Ji, and S. Ananiadou, "Emotion fusion for mental illness detection from social media: A survey," *Inf. Fusion*, vol. 92, pp. 231–246, Apr. 2023, doi: 10.1016/j.inffus.2022.11.031.

[11] Z. Liu, T. Zhang, K. Yang, P. Thompson, Z. Yu, and S. Ananiadou, "Emotion detection for misinformation: A review," *Inf. Fusion*, vol. 107, Jul. 2024, Art. no. 102300, doi: 10.1016/j.inffus.2024.102300.

[12] Y. Ma, K. L. Nguyen, F. Z. Xing, and E. Cambria, "A survey on empathetic dialogue systems," *Inf. Fusion*, vol. 64, pp. 50–70, Dec. 2020, doi: 10.1016/j.inffus.2020.06.011.

[13] H. Xie, S. Feng, D. Wang, and Y. Zhang, "A novel attention based CNN model for emotion intensity prediction," in *Natural Language Processing and Chinese Computing* (Lecture Notes in Computer Science), vol. 11108, M. Zhang, V. Ng, D. Zhao, S. Li, and H. Zan, Eds., Cham, Switzerland: Springer, 2018, pp. 365–377, doi: 10.1007/978-3-319-99495-6_31.

[14] M. S. Akhtar, A. Ekbal, and E. Cambria, "How intense are you? Predicting intensities of emotions and sentiments using stacked ensemble [Application Notes]," *IEEE Comput. Intell. Mag.*, vol. 15, no. 1, pp. 64–75, Feb. 2020, doi: 10.1109/MCI.2019.2954667.

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North*. Minneapolis, MN, USA: Association for Computational Linguistics, 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423.

[16] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.

[17] A. Bello, S.-C. Ng, and M.-F. Leung, "A BERT framework to sentiment analysis of tweets," *Sensors*, vol. 23, no. 1, p. 506, Jan. 2023, doi: 10.3390/s23010506.
[18] W. Liao, B. Zeng, X. Yin, and P. Wei, "An improved aspect-category sentiment analysis model for text sentiment analysis based on RoBERTa," *Int. J. Speech Technol.*, vol. 51, no. 6, pp. 3522–3533, Jun. 2021, doi: 10.1007/s10489-020-01964-1.
[19] D. Yin, T. Meng, and K.-W. Chang, "SentiBERT: A transferable transformer-based architecture for compositional sentiment semantics," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds., Jul. 2020, pp. 3695–3706, doi: 10.18653/v1/2020.acl-main.341.
[20] H. Xu, B. Liu, L. Shu, and P. Yu, "BERT post-training for review reading comprehension and aspect-based sentiment analysis," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, vol. 1, J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, MN, USA: Association for Computational Linguistics, Jun. 2019, pp. 2324–2335, doi: 10.18653/v1/N19-1242.
[21] H. Xu, B. Liu, L. Shu, and P. Yu, "DomBERT: Domain-oriented language model for aspect-based sentiment analysis," in *Proc. Findings Assoc. Comput. Linguistics (EMNLP)*, 2020, pp. 1725–1731, doi: 10.18653/v1/2020.findings-emnlp.156.
[22] B. Zhang, H. Yang, T. Zhou, M. A. Babar, and X.-Y. Liu, "Enhancing financial sentiment analysis via retrieval augmented large language models," in *Proc. 4th ACM Int. Conf. AI Finance*. Brooklyn, NY, USA: ACM, Nov. 2023, pp. 349–356, doi: 10.1145/3604237.3626866.
[23] L. Zhang, H.-L. Huang, Y. Yu, and D. Moldovan, "Affect in tweets: A transfer learning approach," in *Proc. 12th Lang. Resour. Eval. Conf.*, N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, Eds. Marseille, France: European Language Resources Association, 2020, pp. 1511–1516. Accessed: Apr. 17, 2024. [Online]. Available: https://aclanthology.org/2020.lrec-1.188
[24] Y. Zhang, M. Wang, Y. Wu, P. Tiwari, Q. Li, B. Wang, and J. Qin, "DialogueLLM: Context and emotion knowledge-tuned large language models for emotion recognition in conversations," 2023, *arXiv:2310.11374*.
[25] W. Stigall, M. A. Al Hafiz Khan, D. Attota, F. Nweke, and Y. Pei, "Large language models performance comparison of emotion and sentiment classification," in *Proc. ACM Southeast Conf. ZZZ*. Marietta, GA, USA: ACM, Apr. 2024, pp. 60–68, doi: 10.1145/3603287.3651183.
[26] S. Lei, G. Dong, X. Wang, K. Wang, and S. Wang, "InstructERC: Reforming emotion recognition in conversation with a retrieval multi-task LLMs framework," 2023, *arXiv:2309.11911*.
[27] Z. Liu, K. Yang, T. Zhang, Q. Xie, and S. Ananiadou, "EmoLLMs: A series of emotional large language models and annotation tools for comprehensive affective analysis," 2024, *arXiv:2401.08508*.
[28] Q. Qian, B. Tian, M. Huang, Y. Liu, X. Zhu, and X. Zhu, "Learning tag embeddings and tag-specific composition functions in recursive neural network," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics, 7th Int. Joint Conf. Natural Lang. Process.* Beijing, China: Association for Computational Linguistics, Jul. 2015, pp. 1365–1374, doi: 10.3115/v1/p15-1132.
[29] Q. Qian, M. Huang, J. Lei, and X. Zhu, "Linguistically regularized LSTM for sentiment classification," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1. Vancouver, BC, Canada: Association for Computational Linguistics, Jul. 2017, pp. 1679–1689, doi: 10.18653/v1/P17-1154.
[30] M. Huang, Q. Qian, and X. Zhu, "Encoding syntactic knowledge in neural networks for sentiment classification," *ACM Trans. Inf. Syst.*, vol. 35, no. 3, pp. 1–27, Jul. 2017, doi: 10.1145/3052770.
[31] L. He, T. Yin, and K. Zheng, "They may not work! An evaluation of eleven sentiment analysis tools on seven social media datasets," *J. Biomed. Informat.*, vol. 132, Aug. 2022, Art. no. 104142, doi: 10.1016/j.jbi.2022.104142.
[32] C. Hutto and E. Gilbert, "VADER: A parsimonious rule-based model for sentiment analysis of social media text," *Proc. Int. AAAI Conf. Web Social Media*, vol. 8, no. 1, pp. 216–225, May 2014, doi: 10.1609/icwsm.v8i1.14550.
[33] K. A. Govindasamy and N. Palanichamy, "Depression detection using machine learning techniques on Twitter data," in *Proc. 5th Int. Conf. Intell. Comput. Control Syst. (ICICCS)*. Madurai, India: IEEE, May 2021, pp. 960–966, doi: 10.1109/ICICCS51141.2021.9432203.
[34] J. Y.-L. Chan, K. T. Bea, S. M. H. Leow, S. W. Phoong, and W. K. Cheng, "State of the art: A review of sentiment analysis based on sequential transfer learning," *Artif. Intell. Rev.*, vol. 56, no. 1, pp. 749–780, Jan. 2023, doi: 10.1007/s10462-022-10183-8.
[35] D. Dong, F. Lin, G. Li, and B. Liu, "Sentiment-aware fake news detection on social media with hypergraph attention networks," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*. Prague, Czech Republic: IEEE, Oct. 2022, pp. 2174–2180, doi: 10.1109/SMC53654.2022.9945398.
[36] P. Ke, H. Ji, S. Liu, X. Zhu, and M. Huang, "SentiLARE: Sentiment-aware language representation learning with linguistic knowledge," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Nov. 2020, pp. 6975–6988, doi: 10.18653/v1/2020.emnlp-main.567.
[37] S. Mohammad, F. Bravo-Marquez, M. Salameh, and S. Kiritchenko, "SemEval-2018 task 1: Affect in tweets," in *Proc. 12th Int. Workshop Semantic Eval.* New Orleans, LA, USA: Association for Computational Linguistics, 2018, pp. 1–17, doi: 10.18653/v1/s18-1001.
[38] S. A. Qureshi, G. Dias, M. Hasanuzzaman, and S. Saha, "Improving depression level estimation by concurrently learning emotion intensity," *IEEE Comput. Intell. Mag.*, vol. 15, no. 3, pp. 47–59, Aug. 2020, doi: 10.1109/MCI.2020.2998234.
[39] S. Consoli, L. Barbaglia, and S. Manzan, "Fine-grained, aspect-based sentiment analysis on economic and financial lexicon," *Knowl.-Based Syst.*, vol. 247, Jul. 2022, Art. no. 108781, doi: 10.1016/j.knosys.2022.108781.
[40] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "LLaMA: Open and efficient foundation language models," 2023, *arXiv:2302.13971*.
[41] H. Touvron et al., "Llama 2: Open foundation and fine-tuned chat models," 2023, *arXiv:2307.09288*.
[42] Z. Du, Y. Qian, X. Liu, M. Ding, J. Qiu, Z. Yang, and J. Tang, "GLM: General language model pretraining with autoregressive blank infilling," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 320–335, doi: 10.18653/v1/2022.acl-long.26.
[43] A. Zeng, X. Liu, Z. Du, Z. Wang, H. Lai, M. Ding, Z. Yang, Y. Xu, W. Zheng, X. Xia, W. Lam Tam, Z. Ma, Y. Xue, J. Zhai, W. Chen, P. Zhang, Y. Dong, and J. Tang, "GLM-130B: An open bilingual pre-trained model," 2022, *arXiv:2210.02414*.
[44] A. Yang et al., "Baichuan 2: Open large-scale language models," 2023, *arXiv:2309.10305*.
[45] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. Victoria Lin, T. Mihaylov, M. Ott, S. Shleifer, K. Shuster, D. Simig, P. S. Koura, A. Sridhar, T. Wang, and L. Zettlemoyer, "OPT: Open pre-trained transformer language models," 2022, *arXiv:2205.01068*.
[46] G. Team et al., "Gemma: Open models based on Gemini research and technology," 2024, *arXiv:2403.08295*.
[47] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. de Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for NLP," in *Proc. 36th Int. Conf. Mach. Learn.*, vol. 97, K. Chaudhuri and R. Salakhutdinov, Eds., Jun. 2019, pp. 2790–2799. [Online]. Available: https://proceedings.mlr.press/v97/houlsby19a.html
[48] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *Proc. ICLR*, 2022, pp. 1–13. [Online]. Available: https://openreview.net/forum?id=nZeVKeeFYf9
[49] S. Mangrulkar, S. Gugger, L. Debut, Y. Belkada, S. Paul, and B. Bossan. (2022). *PEFT: State-of-the-art Parameter-efficient Fine-tuning Methods*. [Online]. Available: https://github.com/huggingface/peft
[50] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient finetuning of quantized LLMs," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 2023, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds. Red Hook, NY, USA: Curran Associates, 2023, pp. 10088–10115. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/1feb87871436031bdc0f2beaa62a049b-Paper-Conference.pdf
[51] S. Diao, R. Pan, H. Dong, K. S. Shum, J. Zhang, W. Xiong, and T. Zhang, "LMFlow: An extensible toolkit for finetuning and inference of large foundation models," 2023, *arXiv:2306.12420*.
[52] Z. Zheng, L. Liao, Y. Deng, and L. Nie, "Building emotional support chatbots in the era of LLMs," 2023, *arXiv:2308.11584*.
[53] E. Cambria and A. Hussain, "SenticNet," in *Sentic Computing*. Cham, Switzerland: Springer, 2015, pp. 23–71, doi: 10.1007/978-3-319-23654-4_2.
[54] A. Karimi, L. Rossi, and A. Prati, "AEDA: An easier data augmentation technique for text classification," in *Proc. Findings Assoc. Comput. Linguistics (EMNLP)*, Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021, pp. 2748–2754, doi: 10.18653/v1/2021.findings-emnlp.234.

[55] S. Rosenthal, N. Farra, and P. Nakov, "SemEval-2017 task 4: Sentiment analysis in Twitter," in *Proc. 11th Int. Workshop Semantic Eval. (SemEval)*, S. Bethard, M. Carpuat, M. Apidianaki, S. M. Mohammad, D. Cer, and D. Jurgens, Eds. Vancouver, BC, Canada: Association for Computational Linguistics, Aug. 2017, pp. 502–518, doi: 10.18653/v1/s17-2088.
[56] J. Yang and Y. Zhang, "NCRF++: An open-source neural sequence labeling toolkit," in *Proc. ACL Syst. Demonstrations*. Melbourne, VIC, Australia: Association for Computational Linguistics, Jul. 2018, pp. 74–79, doi: 10.18653/v1/P18-4013.

DIEFAN LIN is currently pursuing the degree with Beijing Normal University. He is a Research Assistant with the Institute of Artificial Intelligence and Future Networks, part of the Institute for Advanced Study of Natural Sciences, Beijing Normal University. He leads one research project and participates in another funded by the Science and Technology Innovation Strategy Foundation of Guangdong Province. His research interests include large language models, natural language processing, and sentiment analysis.

YI WEN was born in Shenzhen, Guangdong, China, in 2003. He is currently pursuing the bachelor's degree in data science and big data technology with the Faculty of Arts and Sciences, Beijing Normal University. He is actively preparing to apply to the master's degree with The Hong Kong University. His research interests include deeply rooted in the fields of mathematics, statistics, and machine learning.

WEISHI WANG is currently pursuing the bachelor's degree in artificial intelligence with Beijing Normal University. He is preparing to apply for the Communication Engineering Program at Nanyang Technological University. He was a Research Assistant in the Project "Perception Behavior Experiment Assisted Teaching System Based on Large Models." His research interests include wireless communication and natural language processing.

YAN SU is currently pursuing the bachelor's degree in artificial intelligence with Beijing Normal University. Her research interests include natural language processing and pattern recognition. She is also exploring the intersection of artificial intelligence and biology.