基于混合专家的可扩展情感分析模型

陈 千 [1, 2] 胡梦强 [1] 郭 鑫 [1] 王素格 [1, 2]

摘 要 情感分析作为自然语言处理领域的核心任务之一, 面临着精准捕捉细粒度情感特征以及提升模型可解释性的双重挑战. 为此, 提出一种基于混合专家模型的可扩展情感分析框架, 通过将门控机制融入专家内部, 设计可在任意预训练语言模型中扩展的混合专家模块. 该框架旨在以可控的计算开销扩展模型容量, 促进细粒度条件计算和专家专业化. 在三个典型情感分析数据集上的综合实验表明, 与基线模型相比, 本方法在关键指标上均取得显著提升, 尤其在处理复杂多分类任务时, 其性能已达到甚至超过主流参数高效微调大语言模型的水平. 更重要的是, 得益于稀疏激活机制, 模型在保持高性能的同时, 展现出卓越的推理效率. 通过对专家激活模式和输出表征的深入分析, 清晰地观察到不同专家针对特定语义模式形成功能专精. 这为模型决策提供直观且有力的可解释性证据, 验证该框架在构建高效、高性能且可信赖的情感分析系统中的巨大潜力.

关键词 情感分析; 混合专家; 可解释性; 细粒度特征捕捉; 可扩展性

引用格式 陈千, 胡梦强, 郭鑫, 王素格. 基于混合专家的可扩展情感分析模型. 自动化学报, 2026, **52**(4): 749−764

DOI 10.16383/j.aas.c250366 **CSTR** 32138.14.j.aas.c250366

Scalable Sentiment Analysis Model Based on Mixture of Experts

CHEN Qian[1, 2] HU Meng-Qiang[1] GUO Xin[1] WANG Su-Ge[1, 2]

Abstract As one of the core tasks in natural language processing, sentiment analysis faces dual challenges: Accurately capturing fine-grained emotional features and enhancing model interpretability. To address these issues, we propose a scalable sentiment analysis framework based on a mixture of experts (MoE) model. By integrating a gating mechanism into the expert modules, we design a mixture of experts module that can be extended to any pre-trained language model. The framework aims to expand model capacity with controllable computational overhead, thereby enabling fine-grained conditional computation and expert specialization. Comprehensive experiments on three representative sentiment analysis datasets demonstrate that, compared with baseline models, our approach achieves significant improvements across key metrics. Notably, when handling complex multi-classification tasks, its performance rivals or even surpasses mainstream large language models that have undergone parameter-efficient fine-tuning. More importantly, benefiting from the sparse activation mechanism, the model maintains high performance while exhibiting exceptional inference efficiency. Through an in-depth analysis of expert activation patterns and output representations, we clearly observe that different experts develop functional specialization toward specific semantic patterns, providing intuitive and strong interpretability evidence for model decision-making. These findings validate the great potential of the proposed framework in building efficient, high-performance, and trustworthy sentiment analysis systems.

Keywords sentiment analysis; mixture of experts; interpretability; fine-grained feature capture; scalability

Citation Chen Qian, Hu Meng-Qiang, Guo Xin, Wang Su-Ge. Scalable sentiment analysis model based on mixture of experts. *Acta Automatica Sinica*, 2026, **52**(4): 749−764

收稿日期 2025-08-01 录用日期 2025-12-31
Manuscript received August 1, 2025; accepted December 31, 2025
国家自然科学基金联合重点项目 (U24A20335), 国家自然科学基金 (6237073346) 资助
Supported by National Natural Science Foundation of China Joint Key Project (U24A20335) and National Natural Science Foundation of China (6237073346)
本文责任编委 陶建华
Recommended by Associate Editor TAO Jian-Hua
1. 山西大学计算机与信息技术学院 太原 030006 2. 山西大学计算智能与中文信息处理教育部重点实验室 太原 030006
1. School of Computer and Information Technology, Shanxi University, Taiyuan 030006 2. Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan 030006

情感分析是自然语言处理领域最热门的研究方向之一, 其主要目标是从文本数据中系统地识别、提取并量化主观信息, 如观点、情感和态度等[1]. 随着用户生成数据的爆炸式增长, 从复杂的商业智能系统到实时的社交媒体监控, 情感分析在舆情分析[2]、商业决策[3]、人机交互[4] 等领域得到越来越广泛的应用, 具有重大研究价值.

传统基于规则和机器学习的方法 (如 SVM (support vector machine)、朴素贝叶斯、随机森林等) 虽然简单, 但在识别否定、隐式情感和反讽等复

杂情感时表现不佳[5−11]. 深度学习技术 (如 CNN (convolutional neural network)、LSTM (long short-term memory)) 因自动化特征提取和较高准确性等优势被广泛应用于情感分析领域, 然而细粒度情绪捕获能力不足和海量高质量标注数据稀缺是其固有缺陷[12]. 基于 Transformer 架构的预训练语言模型 (pre-trained language model, PLM) 进一步催生预训练 + 微调范式的诞生 (如 RoBERTa、T5、GPT2), 为识别文本中复杂的语义和情感特征提供强大而灵活的骨干. 然而 PLM 通常采用密集型网络结构, 要求对模型中数量巨大的参数进行全激活, 导致高计算成本和低推理延迟问题, 严重限制了模型在资源受限场景下的应用[13−14].

近年来以 GPT-4、Grok、DeepSeek、LLaMA 等为代表的大型语言模型 (large language model, LLM) 在包括通用情感分析的广泛 NLP 任务中展现出卓越的性能. 然而在理解更细微情感现象和结构化情感信息的复杂任务中, LLM 表现出较大性能差异, 甚至不如在特定领域中专门训练的小型语言模型[13]. LLM 庞大的参数规模使得其在特定领域存在训练成本高、推理计算资源高、实际部署应用成本高、模型可解释性严重不足等缺点. LLM 对情感的嵌入表示编码的具体机制仍未被充分探索, 这进一步限制了部分闭源 LLM 的优化空间和情感可解释性. 情感分析领域前沿任务促使当前研究从对 LLMs 的持续追求转向小型且高效模型、参数高效微调 (parameter-efficient fine-tuning, PEFT) 及基于适配器的方法[15−16].

MoE (mixture of experts) 是近年来提出的一种通过多个专家网络并行处理、门控机制动态组合其输出的模型架构[17]. 单一的、整体的模型往往难以应对所有语言现象, 而情感分析任务的内在复杂性和多面性使其天然适合采用 MoE 的计算范式. MoE 选择性激活专家的能力, 为解决这一问题提供了一种强大的机制. 如图 1 所示, 不同的专家可以被训练或鼓励专门处理情感的不同方面: 第一个专家可能擅长识别隐式情感, 第二个擅长检测否定情感, 第三个擅长理解反讽语气. 通过让每个专家专注于输入数据的不同方面, 模型内部的决策路径可以变得更加可追溯和易于解释. 此外, 基于稀疏门控的专家路由策略使得在推理过程中仅激活少量几个擅长处理当前任务的专家模块, 从而极大降低计算成本. 这对于大规模情感分析任务非常有益, 使得模型集成更大规模的参数成为可能, 进而产生更丰富的语义表示.

然而现有 MoE 的理论研究主要聚焦于门控路

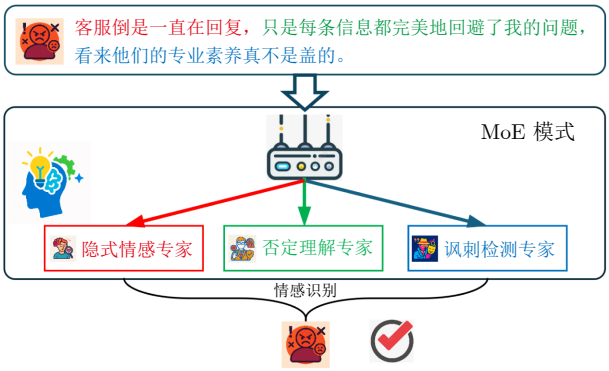


图 1 MoE 模式在情感分析中的示例

Fig. 1 Example of MoE model for sentiment analysis

由策略 (从简单路由到 token 级、模态级、任务级以及复杂动态算法路由)、负载均衡、训练效率、降低计算资源和通信开销的相关研究[18], 对专家模型内部的设计关注较少, 且由于 MoE 专家结构与前馈网络 (feedforward network, FFN) 差异较大, 直接替换可能导致预训练知识丢失[19−20]. 此外, 现有的 MoE 架构大多将专家视为静态的前馈网络, 其条件化计算完全依赖于顶层的路由网络. 这种设计虽然降低了计算量, 但一旦输入被分配给某个专家, 其内部处理流程是固定的, 这限制了专家处理多样化特征的灵活性. 与传统 MoE 将专家视为静态 FFN 不同, 本工作在情感分析任务中探索专家内部的门控机制, 提出一种两阶段条件计算 MoE 框架. 第一阶段, 通过稀疏门控路由网络选择宏观层面的专家; 第二阶段, 在每个专家内部嵌入一个独立的门控单元, 实现专家内部的微观条件计算. 这意味着每个专家都能根据接收到的特征动态调整其内部信息流, 从而学习到更具区分度与专精化的表征. 这种设计不仅提升模型的细粒度情感捕捉能力, 更重要的是, 其还能通过分析输入–输出在专家簇中的分布差异, 为模型的可解释性提供新的维度, 有效回应 LLM 时代对高效、透明 AI 的需求.

本文的主要贡献包括: 1) 提出一种在性能与效率之间取得卓越平衡的可扩展 MoE 模型框架. 实验证明, 该框架在显著提升情感分析性能 (在 Tweet-Eval 等复杂任务上超越了多种微调后的 LLM) 的同时, 保持了远超基线模型的推理吞吐量和更低的计算量, 为解决大模型在实际应用中部署和推理成本高昂的问题提供了新思路. 2) 从多个维度 (专家激活热图、输出表征聚类、具体案例分析) 深入地探究模型的可解释性. 提供专家功能专精化的直接可视化证据, 清晰地揭示模型处理混合情感、反讽和双重否定等复杂语言现象时的内部决策路径. 3) 对所提出的 MoE 框架进行全面深入的实证评估, 不

仅包含了与多种参数高效微调方法的横向对比, 还涵盖了与多种主流大语言模型 (LLMs) 在零样本和微调设置下的性能及资源开销对比, 明确了本方法在当前技术格局中的独特优势和适用场景.

1　相关工作

近年来情感分析领域由最初的基于词典[5−7] 和传统机器学习范式[8−9, 11−13], 逐渐过渡到基于深度学习的方法[10], 并进入到以预训练语言模型 (PLM) 为核心的时代. 近年来, PLM 的规模不断扩大, 催生了具有更强通用能力的 LLM[13−14, 17].

基于深度学习的情感分析方法在过去十年里得到显著普及和广泛应用, 循环神经网络 (LSTM、GRU) 和图神经网络能够有效捕捉文本的序列信息和长距离依赖. 卷积神经网络 CNN 则擅长提取局部特征. 随后, 注意力机制的引入使得模型能够关注文本中对情感判断更重要的部分[21]. 然而上述模型对细粒度情感特征的区分能力有限. Transformer 模型的提出让情感分析领域进入基于预训练 + 微调范式时代. 它以自注意力机制为核心, 使其能够并行处理并捕捉文本中的长距离依赖关系, 这与传统的循环和卷积结构形成了鲜明对比. 如基于 BERT 和基于 RoBERTa 的情感分析方法能够自动学习文本语义特征, 避免繁琐的人工特征工程, 显著提升了情感分析的性能[22−23]. 大语言模型在包括通用情感分类在内的广泛 NLP 任务中展现出卓越的能力和强大的性能[13], 然而 LLM 表示情感编码的具体机制仍未被充分探索, 这限制了情感分析性能的优化空间和可解释性.

近年来, MoE 作为一种有效方法, 在大幅提升模型容量的同时, 最大限度减少计算开销, 逐渐成为应对现代 AI 模型扩展性挑战的核心范式, 在学术界和工业界得到广泛关注[2]. MoE 会动态地选择并激活一个相关的参数子集, 稀疏激活机制是其效率的关键所在, 从而在不按比例增加计算成本的情况下显著扩展模型容量[24]. Shazeer 等[25] 于 2017 年提出一种稀疏门控路由机制, 以替换 Transformer 块中传统的密集型前馈神经网络层. 之后, MoE 在 LLM 领域取得了显著成功. Google 的 Switch Transformer 首次将 MoE 应用于万亿参数级别的模型, 展示了其在超大规模模型训练中的潜力[26]. DeepMind 的 GLaM[27] 和 Google Brain 的 LaMDA[28] 也采用了 MoE 架构. 这些模型在保持计算效率的同时, 实现了模型规模的扩展和性能的提升. 尽管 MoE 模型在性能和效率方面表现出色, 但其可解释性仍然是一个挑战. 理解门控网络如何选择专家以及每个专家具体学习了什么, 对于模型的调试和改进至关重要.

MoE 架构的一个持续存在的挑战在于确保专家真正学习到独特、非冗余的知识. 常见的辅助负载均衡损失虽然对于防止专家崩溃 (即少数专家主导) 至关重要, 但有时会导致专家重叠或同质化表示, 从而阻碍真正的专精化. 为了分析专家行为和专精化, 已有研究提出各种方法[29], 包括可视化注意力图、激活模式以及应用可解释人工智能技术, 例如 SHAP 值. 最近的研究引入了正交性损失和方差损失等特定目标, 旨在积极促进独特的专家表示并鼓励更具区分性的路由决策. 这些技术旨在明确减少专家重叠并增加路由分数方差, 从而增强专精化. 现有研究主要专注于 MoE 的外在调整, 很少关注单个专家对象, 且较少关注 MoE 在情感分析中的应用以及如何通过专家分工提升模型的专业化程度和可解释性. 与现有工作不同的是, 本文将 MoE 应用于情感分析任务, 在每个专家内部设计和嵌入一个门控单元, 这意味着专家不是静态的 FFN, 而是在被主路由网络选择后, 对它们接收到的输入执行某种形式的条件计算, 达到基于输入特征动态调整其内部处理的目的. 这种细粒度的内部门控机制可以使每个专家内部实现更深层次、更细致的专精化, 从而有效地应对这些挑战.

2　模型介绍

本文提出一种可扩展的 MoE 模块框架, 如图 2 所示, 该框架以 RoBERTa 为示例, 编码器一共有 L 层, 模型的前半部分 (前 M 层) 与标准 Transformer 块一样, 目的是利用初始预训练知识加速训练过

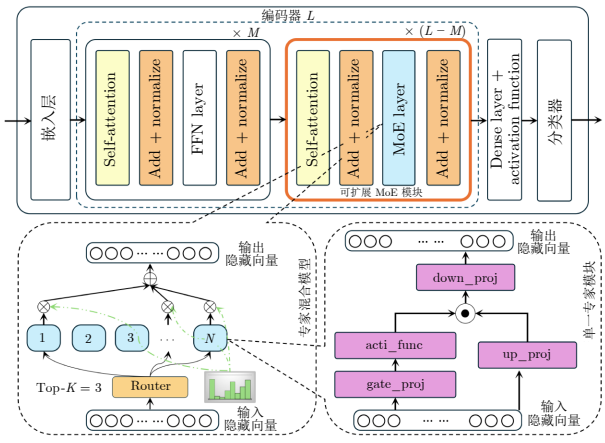


图 2　基于 RoBERTa 的可扩展 MoE 框架

Fig. 2　An extensible MoE framework based on RoBERTa

程, 同时降低训练成本. 而在 $L-M$ 层中将 Transformer 块中的 FFN 层替换为自定义的 MoE layer, 目的是发挥专家网络的特性, 促进更细粒度的特征提取, MoE layer 的设计细节将在第 2.2.2 节介绍. 采用这种设计的目的在于在尽可能不改变预训练模型架构的基础上最大限度提高模型的性能, 可以根据资源限制控制 M 的值来决定训练成本 (如果 M 值为 0, 意味着要将预训练模型中所有层都进行替换, 从头开始训练).

2.1 MoE-RoBERTa 模型架构

本节将自定义的 MoE 层集成到 RoBERTa 模型中, 构建了 MoE-RoBERTa 模型. 具体而言, 选择替换 RoBERTa 模型后几层的 FFN 为设计的 MoE 层. 这种替换策略基于以下考虑:

1) 深层语义处理. RoBERTa 模型的前几层主要负责提取文本的低级和中级语义特征, 而模型的后几层则更侧重于处理高级语义和任务相关的特征. 在这些深层引入 MoE, 可以使模型在处理复杂情感模式时, 动态地调用不同的专家来捕捉细微的情感差异.

2) 计算效率与性能平衡. 替换所有 FFN 层可能会带来过高的计算开销, 而仅替换部分深层 FFN 可以在保持计算效率的同时, 有效利用 MoE 的优势提升模型性能.

MoE-RoBERTa 的整体架构保持了 RoBERTa 的 Transformer 编码器堆叠结构, 但在指定的层中, 将原有的 FFN 模块替换为专家混合模型. 模型的输入首先经过 RoBERTa 的嵌入层, 然后通过多层编码器, 在 RoBERTa 的输出之上添加一个分类器, 用于将文本表示映射到情感类别.

2.2 专家混合模型

为了提升传统 FFN 的计算效率和模型容量, 并引入专家专精的特性, 本文设计了一种自定义的 MoE 层来替换 RoBERTa 模型中的 FFN. 本文提出的 MoE 层由一个路由网络和多个专家模块组成. 每个专家模块本质上是一个优化的 FFN.

2.2.1 路由网络

路由网络, 是 MoE 系统的大脑, 负责动态地将输入 token 分配给最相关的专家. 其智能设计对于实现有效的专家专业化和保持专家利用率平衡至关重要. 在稀疏门控的混合专家模型中, 路由函数 $g(\boldsymbol{x})$ 的目标是为输入 $\boldsymbol{x}\in\mathbf{R}^d$ 分配多个专家的权重. 本文采用一种引入噪声的门控机制以增强专家的多样性选择能力, 其计算过程如下:

$$\begin{cases}\boldsymbol{s}=\boldsymbol{x}\boldsymbol{W}_{\text{gate}}+\boldsymbol{b}_{\text{gate}}\\ \epsilon_i\sim \text{N}(0,\,\sigma^2),\quad i=1,\,2,\,\cdots,\,E\\ \tilde{\boldsymbol{s}}=\boldsymbol{s}+\boldsymbol{\epsilon}\\ g(\boldsymbol{x})=\text{softmax}(\tilde{\boldsymbol{s}})\end{cases}\tag{1}$$

其中, $\boldsymbol{W}_{\text{gate}}\in\mathbf{R}^{d\times E}$ 和 $\boldsymbol{b}_{\text{gate}}\in\mathbf{R}^E$ 为门控网络的可学习参数; E 表示专家数量; $\boldsymbol{\epsilon}\in\mathbf{R}^E$ 为与每个专家对应的独立高斯噪声项, 其各维服从均值为 0、方差为 σ^2 的正态分布. 通过引入噪声 $\boldsymbol{\epsilon}$, 可有效提升训练初期专家选择的多样性. 推理阶段则不添加噪声以确保模型输出的稳定性.

2.2.2 专家网络

路由网络负责在宏观层面将 token 分配给专家, 通常会结合负载均衡机制以确保每个专家被均匀利用. 然而, 如果它们被迫处理各种各样的输入, 这可能无意中导致专家过度泛化. 因此有必要设计一种机制使得即使一个专家由于外部负载均衡而接收到一些多样化的 token, 其仍然可以通过学习动态地关注这些 token 的不同方面. 每个专家模块 E_i 是一个独立的前馈模块, 旨在专注于对输入特征的特定子空间建模. 不同于传统的 FFN, 每个专家 E_i 的结构如下:

$$E_i(\boldsymbol{x})=\text{down\_proj}\Big(\text{acti\_func}(\text{gate\_proj}(\boldsymbol{x}))\odot \text{up\_proj}(\boldsymbol{x})\Big)\tag{2}$$

其中:

- $\boldsymbol{x}$ 表示 MoE 层的输入, 通常为来自前一多头自注意力层的输出.
- gate_proj(·) 是一层线性变换, 用于生成门控信号. 此操作在每个专家内部创建了一个乘法门控机制. 这种内部门控允许每个专家根据接收到的输入特定特征动态地强调或抑制其内部表示.
- $\odot$ 表示逐元素相乘, 门控信号与 up_proj($\boldsymbol{x}$) 的结果逐元素相乘, 实现对特征维度的动态选择.
- acti_func(·) 表示激活函数, 用于引入非线性.
- up_proj(·) 为线性变换层, 用于将输入映射到更高维空间.
- down_proj(·) 为线性变换层, 用于将高维表示压缩回原始维度.

虽然路由网络有效地决定了哪个专家处理给定的 token, 但这种内部门控使选定的专家能够以更优的适应性和选择性处理路由网络分配的 token. 这种架构设计可以通过使专家学习独特的内部转换来促进更深入、更细致的专精化, 从而提取出更丰

富、更解耦的特征表示. 这种外部路由和内部门控之间的细致交互, 有望为 MoE 架构中长期存在的负载均衡与专精化之间的权衡提供一个新颖的解决方案.

2.2.3　融合网络

融合网络是根据路由网络的输出概率对被选中的 Top-K 专家输出进行加权求和. 具体来说, 输入 $\boldsymbol{x}$ 的前向计算流程如下所示:

1) 输入 $\boldsymbol{x}$ 经由路由网络生成各专家对应的权重 g_i;

2) 根据权重大小选择 Top-K 个专家 $\{E_{i_1}, \cdots, E_{i_K}\}$ 进行激活;

3) 每个被激活的专家 E_i 处理输入 $\boldsymbol{x}$, 产生对应的输出 $E_i(\boldsymbol{x})$;

4) MoE 层的最终输出为被选中专家输出的加权和, 表示为:

$$\mathrm{MoE}(\boldsymbol{x}) = \sum_{i \in \mathcal{S}(\boldsymbol{x})} g_i \cdot E_i(\boldsymbol{x}) \tag{3}$$

其中, $\mathcal{S}(\boldsymbol{x})$ 表示根据门控权重选择的 Top-K 个专家索引集合, g_i 为专家 E_i 对应的权重.

该融合方式可实现专家行为的稀疏激活与加权整合, 使模型能够动态选择最适合当前输入的专家子网络.

2.3　训练策略

除了标准的交叉熵损失之外, 本文还引入了专家负载均衡损失和 L_2 正则化项. 负载均衡损失旨在鼓励门控网络将输入均匀地分配给所有专家, 确保每个专家都能得到充分的训练. L_2 正则化项旨在避免路由网络输出值过大导致专家分配过于极端, 从而提升模型训练稳定性. 总损失函数定义如下:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \alpha \left(\mathcal{L}_{\text{aux}} + \beta \cdot \mathcal{L}_{\text{z}} \right) \tag{4}$$

其中, 各项定义如下:

$$\mathcal{L}_{\text{CE}} = -\sum_{c=1}^{C} \mathbb{1}_{[y=c]} \ln \hat{y}_c \tag{5}$$

$$\mathcal{L}_{\text{aux}} = E \cdot \sum_{i=1}^{E} l_i \cdot p_i \tag{6}$$

$$\mathcal{L}_{\text{z}} = \frac{1}{B \cdot E} \sum_{b=1}^{B} \sum_{i=1}^{E} z_{b,\,i}^2 \tag{7}$$

符号说明如下:

- $\mathcal{L}_{\text{total}}$ 表示最终的总损失函数;
- $\mathcal{L}_{\text{CE}}$ 表示情感分类任务的交叉熵损失;
- $\hat{y}_c$ 表示模型对类别 c 的预测概率;
- y 为真实标签, C 为类别总数, $\mathbb{1}_{[y=c]}$ 是指示函数, 当 $y = c$ 时取 1, 否则为 0;
- $\mathcal{L}_{\text{aux}}$ 为辅助负载均衡损失;
- E 表示专家总数;
- p_i 表示第 i 个专家被分配到的平均概率, 即 $p_i = \mathrm{E}_{x \,\in\, \text{batch}}[\mathrm{softmax}(\text{router}_{\text{logits}})_i]$;
- l_i 表示第 i 个专家在实际计算中被选中的平均频率;
- $\mathcal{L}_{\text{z}}$ 是 $\text{router}_{\text{logits}}$ 的正则化项, 用于抑制过大的激活值;
- $z_{b,\,i}$ 表示第 b 个样本在第 i 个专家上的 $\text{router}_{\text{logits}}$ 输出;
- B 表示 batch size;
- α 和 β 是超参数, 用于平衡各项损失的相对重要性.

3　实验

3.1　数据集

为全面评估模型在不同情感粒度上的表现, 本文选用了三个具有代表性且被广泛使用的情感分析数据集, 涵盖二分类、多分类等任务场景, 分别为 IMDb、TweetEval Emotion 与 SST-5.

IMDb 是一个电影评论数据集, 广泛用于情感分类研究. 该数据集为二分类任务, 标注了电影评论的情感倾向. 评论内容以自然语言呈现, 语义信息丰富, 适用于评估模型在面向主观文本时的理解能力.

TweetEval Emotion 是 TweetEval benchmark 中的情感识别子任务, 基于 Twitter 上用户发布的推文进行标注. 该任务共有 4 个情感类别, 分别为喜悦、愤怒、悲伤、乐观. 该数据集的语言形式口语化, 结构松散, 挑战性较高, 适合评估模型在社交媒体场景中的鲁棒性.

SST-5 是一个基于电影评论语料的五分类情感数据集, 情感类别包括: 非常负面、负面、中性、正面与非常正面. 与 IMDb 不同的是, SST-5 提供了更细粒度的情感标签, 适合用于测试模型对细微情感差异的识别能力.

表 1 展示了各数据集的训练集、验证集和测试集的样本数量及类别数等统计信息.

3.2　实验设置

3.2.1　模型设置

RoBERTa[30] 是 Facebook AI 在 2019 年提出

表 1 情感分析数据集统计信息
Table 1 Statistics of sentiment analysis datasets

统计项	训练集样本数	验证集样本数	测试集样本数	类别数
IMDb	25 000	—	25 000	2
TweetEval Emotion	3 260	374	1 420	4
SST-5	8 540	1 100	2 210	5

的 BERT 的优化版本. RoBERTa 在 BERT 的基础上进行了多项改进, 包括更多的训练数据和更长的训练时间、使用更大规模的语料库进行更长时间的训练、在训练过程中动态生成掩码、移除下一句预测任务, 这些改进使得 RoBERTa 成为当时 SOTA 的预训练模型之一. 因此在本实验中选用 RoBERTa 作为基线模型.

本实验将 RoBERTa 后两层中的 FFN 替换为 MoE 架构, 具体来说: MoE-RoBERTa 模型中未被替换的部分将使用预训练的 RoBERTa-base 的权重进行初始化, 被替换的部分使用 He 初始化策略对网络权重进行初始化, 其数学表达式如下:

$$W \sim \mathrm{N}\left(0,\ \frac{2}{n_{\mathrm{in}}}\right) \tag{8}$$

其中, W 为需要初始化的权重矩阵, n_{in} 为当前层的输入维度. 即对每一层权重 W 按 $\mathrm{N}(0,\ 2/n_{\mathrm{in}})$ 的正态分布进行采样. 该策略被广泛应用于深层网络中, 能有效缓解梯度消失问题. 同时, 采用 MoE 架构部分替换 FFN 能显著减少全局门控网络和各个专家的初始学习负担, 使它们能够更快地专注于情感特定模式的专精化和学习高效的条件路由, 从而提高了本研究方法的实际可行性和整体效率.

本节选择准确率 (Accuracy)、精确率 (Precision)、召回率 (Recall)、F1 分数 (F1-Score)、混淆矩阵 (Confusion Matrix) 评估模型的性能.

3.2.2 其他实验设置

在模型训练过程中, 本文采用 AdamW 优化器对参数进行更新, 初始学习率设为 2×10^{-5}. 为提升模型的收敛效率与稳定性, 训练过程中引入余弦学习率调度策略, 并设置 10% 的预热步数. 具体地, 训练轮数设置为 5 轮, 总训练步数为总批次数与轮数的乘积; 前 10% 步数执行线性预热, 随后采用余弦退火策略动态调整学习率. 此外, 本文模型采用稀疏门控机制中的 Top-K 路由策略以提升计算效率与参数利用率, 其中 $K = 1$, 即每个样本在每次前向传播中仅激活一个专家模块, 从而构建稀疏的专家路由路径. 实验使用一块 NVIDIA A100 40 GB GPU 进行模型训练.

3.3 实验结果分析

3.3.1 主实验结果分析

为验证 MoE 架构的有效性, 本节旨在通过量化指标, 将 MoE-RoBERTa 模型与基线模型 (RoBERTa) 以及当前前沿的大型语言模型 (LLMs) 在三个典型的情感分析任务上进行性能对比. LLMs 的评估涵盖了零样本 (Zero-shot) 和使用 LoRA 进行参数高效微调两种方式.

具体实验结果如表 2 所示. 首先, 在所有三个数据集上, 本文提出的 MoE-RoBERTa 模型相较于基线 RoBERTa 模型均取得了显著的性能提升. 在 IMDb 数据集上, F1 分数提升了 0.22 个百分点, 该提升很小, 可能由于 IMDb 是二分类任务, 情感区分边界相对清晰, 强大的 RoBERTa 基线模型已能达到很高的性能, MoE 架构精细划分决策边界的优势在这种简单任务上难以完全体现. 但是在更具挑战性的 TweetEval 和 SST-5 数据集上, 性能优势更为明显, F1 分数分别提升了约 1.38 和 3.21 个百分点, 这初步验证了 MoE 架构通过专家分工增强模型对细粒度情感捕捉能力的有效性.

在与大型语言模型的对比中, 观察到以下几点:

1) LLM 微调后的强大性能. 在 IMDb 和 SST-5 这两个数据集上, 经过 LoRA 微调的 LLMs, 特别是 Mistral-7B-Instruct, 展现出了最强的性能. 例如, 在 SST-5 上, 其 F1 分数达到了 0.588 8, 超越了所有其他模型. 这表明 LLMs 凭借其庞大的参数规模和预训练知识, 在进行任务适配后通常能达到很高的性能上限.

2) MoE 架构在特定任务上的竞争力. 值得注意的是, 在四分类的 TweetEval 数据集上, MoE-RoBERTa 模型以 83.27% 的 F1 分数取得了最佳表现, 不仅优于基线, 也超过了包括 Llama-3 和 Mistral-7B 在内的所有经过 LoRA 微调的大型语言模型. 这一结果尤为关键, 它表明本文提出的 MoE 架构在处理类别更多、语境更复杂的社交媒体文本时具有独特的优势, 其专家机制能够有效应对该场景下的情感识别挑战.

3) 零样本方法的局限性. LLMs 在零样本设置下的表现普遍不稳定, 尤其是在 TweetEval 和 SST-5 这两个多分类任务上, 其性能远不及经过微调的模型, 甚至不如基线 RoBERTa 模型. 这说明对于特定领域的细粒度情感分析任务, 直接应用通用大模型是远远不够的.

图 3 给出混淆矩阵的实验结果, 在二分类任务中, 引入 MoE 后, 模型对正类样本的召回率提升: 正确识别的正类样本数量从 11 947 提高到 12 077,

表 2　不同数据集上基线模型与 MoE 增强模型之间的性能比较

Table 2　Performance comparison between baseline and MoE-enhanced models on different datasets

数据集	模型结构	Accuracy	Precision	Recall	F1-Score
IMDb (二分类)	基线	0.9543	0.9543	0.9543	0.9543
	MoE-RoBERTa	0.9565	0.9567	0.9565	0.9565
	Llama-3-8B-Instruct + Zero-shot	0.9333	0.9164	0.9534	0.9346
	Mistral-7B-Instruct + Zero-shot	0.9048	**0.9751**	0.8303	0.8969
	Qwen2.5-7B-Instruct + Zero-shot	0.9424	0.9589	0.9240	0.9411
	Llama-3-8B-Instruct + LoRA	0.9692	0.9660	0.9726	0.9693
	Mistral-7B-Instruct + LoRA	**0.9743**	0.9724	**0.9763**	**0.9743**
	Qwen2.5-7B-Instruct + LoRA	0.9659	0.9641	0.9678	0.9660
TweetEval (四分类)	基线	0.8191	0.8187	0.8191	0.8189
	MoE-RoBERTa	0.8325	**0.8338**	**0.8325**	**0.8327**
	Llama-3-8B-Instruct + Zero-shot	0.7570	0.7846	0.6611	0.6910
	Mistral-7B-Instruct + Zero-shot	0.7720	0.7265	0.7433	0.7296
	Qwen2.5-7B-Instruct + Zero-shot	0.7735	0.7416	0.7226	0.7265
	Llama-3-8B-Instruct + LoRA	0.8248	0.8018	0.7978	0.7986
	Mistral-7B-Instruct + LoRA	**0.8381**	0.8177	0.7902	0.8019
	Qwen2.5-7B-Instruct + LoRA	0.7994	0.7974	0.7878	0.7928
SST-5 (五分类)	基线	0.5452	0.5621	0.5452	0.5464
	MoE-RoBERTa	0.5805	0.5788	0.5805	0.5785
	Llama-3-8B-Instruct + Zero-shot	0.3898	0.2697	0.3821	0.2495
	Mistral-7B-Instruct + Zero-shot	0.4760	0.3922	0.4033	0.3531
	Qwen2.5-7B-Instruct + Zero-shot	0.4841	0.4095	0.4322	0.3952
	Llama-3-8B-Instruct + LoRA	0.5498	0.5568	0.5515	0.5541
	Mistral-7B-Instruct + LoRA	**0.6158**	**0.6148**	**0.5864**	**0.5888**
	Qwen2.5-7B-Instruct + LoRA	0.5398	0.5520	0.5429	0.5474

漏判的正类样本数减少, 说明模型在正类情感上的识别能力增强. 然而, 原本应为负类的样本被错误分类为正类的数量有所上升. 在五分类任务中, MoE 架构在多个情感类别上的预测准确率得到提升, 特别是在极端情感类别中主对角线值增大, 说明架构在细粒度情感区分上具有更强的表达能力. 在四分类任务中, MoE 架构在喜悦等类别上的正确识别数量进一步提升, 同时情感类别之间的混淆程度降低, 表明其在多情感识别任务中具有更强的鲁棒性.

综合来看, 主实验结果充分证明了提出的 MoE-RoBERTa 模型不仅全面超越了其基线模型, 而且在性能上极具竞争力. 尽管经过 LoRA 微调的大型语言模型在某些任务上表现更优, 但考虑到它们巨大的参数规模和随之而来的高昂计算成本 (将在第 3.3.2 节详细分析), MoE 架构在特定复杂任务 (如 TweetEval) 上实现了超越, 展示了其作为一个高效且强大的情感分析解决方案的巨大潜力.

3.3.2　模型效率分析

为全面、多维度地评估 MoE 架构在性能与资源效率方面的综合表现, 本节将其与多种基线模型和前沿方法进行深入比较. 比较对象如下: 1) LLMs, 与主实验相同的模型; 2) 基线模型 (RoBERTa), 对 RoBERTa 进行全量微调 (Full fine-tuning, Full FT); 3) 参数高效微调 (PEFT) 方法, 在 RoBERTa 的基础上, 应用了 BitFit、LoRA 和 P-Tuning 的 PEFT 技术. 本节选用 F1 分数作为核心的模型性能评估指标. 同时, 为衡量模型效率, 引入五个关键的效率指标: 可训练参数量、总参数量、推理吞吐量 (Throughput)、计算量 (GFLOPs) 以及峰值训练显存 (Peak Mem (MB)). 所有实验结果详见表 3.

大型语言模型的性能与代价. 从表 3 中可以清晰地看到 LLMs 通过 LoRA 微调后, 在 IMDb 数据集上展现出了顶尖的性能, F1 分数均超过了 0.96, 显著优于所有基于 RoBERTa 的模型. 这得益于它们庞大的参数规模和强大的预训练知识. 然而, 这种卓越性能的背后是巨大的资源开销. LLMs 的总参数量均在数十亿级别, 其 LoRA 微调所需的峰值训练显存 (11 ~ 19 GB) 远超小型模型. 更重

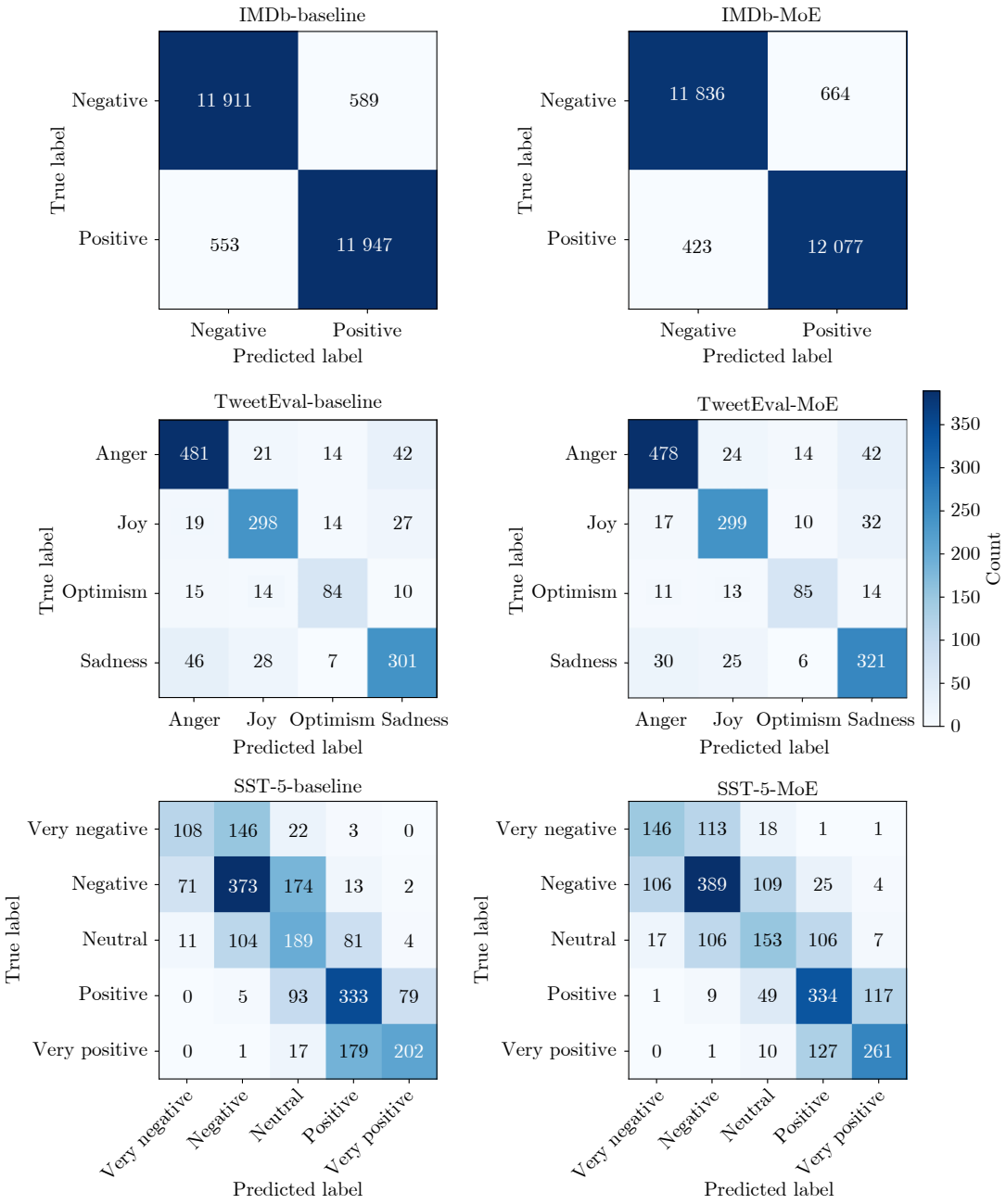


图 3 混淆矩阵图

Fig. 3 Confusion matrix plot

要的是, 它们的推理吞吐量极低 (约 2 ~ 3 samples/s), 这使得它们在需要高并发、低延迟的实际应用场景中面临巨大挑战. Zero-shot 方法虽然无需训练, 推理速度较快, 但其性能相比微调后有明显差距.

基线与参数高效微调 (PEFT) 方法的权衡. 与 RoBERTa 的全量微调 (Full FT) 相比, PEFT 方法 (特别是 BitFit 和 LoRA) 在效率上表现出显著优势. 它们仅通过训练极少数 (0.10 ~ 0.89 M) 的参数, 大幅降低了峰值训练显存 (约 1.3 ~ 1.5 GB), 几乎只有全量微调的一半. 然而, 这种效率的提升往往伴随着性能的牺牲. 在所有三个数据集上, 这些 PEFT 方法的 F1 分数均未能超越全量微调, 其中 P-Tuning 和 BitFit 的性能下降尤为明显. 这表明, 虽然 PEFT 能够有效节约训练资源, 但在追求极致性能时, 它们的能力受到了限制.

MoE-RoBERTa: 性能与效率的卓越平衡. 本文提出的 MoE-RoBERTa 模型在本次对比分析中表现突出, 成功地在性能和效率之间取得了卓越的平衡.

表 3 跨数据集和模型的性能与资源比较

Table 3 Performance and resource comparison across datasets and models

数据集	基座模型	方法	F1-Score	Train (M)	Total (M)	Throughput (samples/s)	GFLOPs	Peak Mem (MB)
IMDb	Llama-3-8B-Instruct	LoRA	0.9693	41.94	4582.54	2.997	4346.11	14955.8
		Zero-shot	0.9346	—	4582.54	9.926	4037.17	—
	Mistral-7B-Instruct	LoRA	0.9743	41.94	3800.31	2.239	4746.59	11248.3
		Zero-shot	0.8969	—	3800.31	10.077	4327.97	—
	Qwen2.5-7B-Instruct	LoRA	0.9660	40.37	4393.34	2.997	3634.03	19273.1
		Zero-shot	0.9411	—	4393.34	8.187	3724.80	—
	RoBERTa	BitFit	0.9036	0.10	124.65	873.470	45.90	1326.3
		Full FT	0.9378	124.65	124.65	796.689	45.90	2702.8
		LoRA	0.9310	0.89	125.53	659.302	46.05	1512.8
		P-Tuning	0.7966	0.61	125.25	857.970	49.69	1439.2
		MoE	0.9426	172.42	172.42	**900.251**	**41.56**	7429.5
TweetEval	Llama-3-8B-Instruct	LoRA	0.8083	41.94	4582.54	2.353	1171.26	11418.3
		Zero-shot	0.6910	—	4582.54	50.868	1159.01	—
	Mistral-7B-Instruct	LoRA	0.8019	41.94	3800.31	2.214	1280.33	11241.9
		Zero-shot	0.7296	—	3800.31	48.024	1266.92	—
	Qwen2.5-7B-Instruct	LoRA	0.8013	40.37	4393.34	3.124	1016.73	14166.8
		Zero-shot	0.7265	—	4393.34	42.518	1010.61	—
	RoBERTa	BitFit	0.1410	0.10	124.65	882.097	45.90	1326.3
		Full FT	0.7962	124.65	124.65	835.678	45.90	2701.5
		LoRA	0.5932	0.89	125.54	628.113	46.05	1512.8
		P-Tuning	0.1410	0.61	125.26	889.742	49.69	1439.3
		MoE	0.8039	200.74	200.74	**890.958**	**42.65**	8152.4
SST-5	Llama-3-8B-Instruct	LoRA	0.5692	41.94	4582.54	2.385	1218.74	12301.7
		Zero-shot	0.2495	—	4582.54	48.552	1225.63	—
	Mistral-7B-Instruct	LoRA	0.5888	41.94	3800.31	2.204	1381.64	11578.2
		Zero-shot	0.3531	—	3800.31	44.868	1378.78	—
	Qwen2.5-7B-Instruct	LoRA	0.5649	40.37	4393.34	3.214	1093.97	15825.3
		Zero-shot	0.3952	—	4393.34	40.558	1061.08	—
	RoBERTa	BitFit	0.0870	0.10	124.65	888.373	45.90	1326.3
		Full FT	0.5432	124.65	124.65	853.694	45.90	2699.3
		LoRA	0.4805	0.89	125.54	651.461	46.05	1512.8
		P-Tuning	0.1385	0.61	125.26	860.703	49.69	1439.3
		MoE	0.5532	172.42	172.42	**910.685**	**41.93**	8404.0

在性能方面, MoE-RoBERTa 在所有三个数据集上的 F1 分数不仅全面超越了基线 RoBERTa 的全量微调, 也显著优于所有被评估的 PEFT 方法. 尤其在 TweetEval 和 SST-5 这两个更具挑战性的多分类任务上, 性能提升尤为显著, 证明了 MoE 架构通过专家分工, 有效提升了模型对细粒度情感的捕捉能力.

在效率方面, MoE 架构的总参数量 (约 172 ~ 200 M) 和可训练参数量相较于全量微调有所增加, 这是因为引入了多个专家网络以扩展模型容量, 从而换取性能的提升. 相应地, 其峰值训练显存 (约 7.4 ~ 8.4 GB) 也高于基线模型. 然而, 最值得关注的是, MoE 架构在推理效率上表现极为亮眼. 其推理吞吐量 (约 900 samples/s) 在所有 RoBERTa 变体中达到最高, 甚至超过了结构更简单的全量微调模型, 这直接得益于其稀疏激活的特性, 即在推理时仅调用部分专家网络. 同时, 其计算量 (GFLOPs) 也是所有 RoBERTa 方法中最低的, 进一步验证了该架构在推理阶段的高效性.

综合分析表明, 本研究提出的 MoE-RoBERTa 模型为情感分析任务提供了一个极具竞争力的解决方案. 它克服了传统全量微调训练成本高和 PEFT

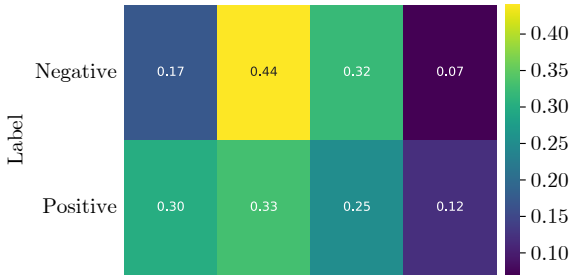
方法性能受限的缺点, 也避开了 LLMs 巨大的部署和推理成本. 本文的模型以一种可控的训练资源开销, 换来了超越基线和 PEFT 方法的性能, 并维持了极高的推理效率. 这有力地证明了将 MoE 融入预训练语言模型顶层的策略, 是在不牺牲性能的前提下, 有效扩展模型容量并兼顾计算效率的一条可行路径.

3.3.3 专家激活模式分析

本实验旨在深入理解在不同情感倾向的输入文本下, MoE 层中的专家是如何被激活的, 以及这种激活模式是否揭示了专家学习到特定模式或知识, 从而增强模型的可解释性.

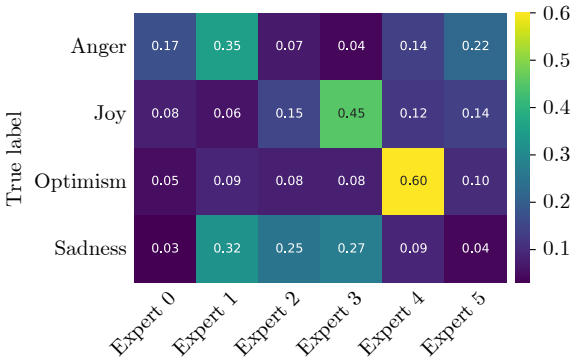
本节抽取测试集中不同情感类别的样本, 并记录这些样本通过 MoE 层时, 不同情感类别样本中的 token 更倾向于激活哪些专家. 通过可视化这些激活分布, 可以观察到不同专家在处理不同类型输入时的活跃程度.

由于越靠近任务层, 模型提取的情感信息就越强烈, 本节只对最后一层的 MoE 进行分析. 在 IMDb 数据集上 (图 4(a)), 负面情感主要激活专家 1 (0.44) 和专家 2 (0.32), 而正面情感更倾向于专家 1 (0.33) 和专家 0 (0.30). 这种差异化激活模式表明专家在学习情感特征时出现了一定程度的功能分化, 但是这种差异没有达到特别明显的程度, 可能是由于一个句子的情感可以由多种语言特征决定 (例如词汇选择、句法结构、否定、反讽等). 专家可能学习了更细粒度的语言特征, 而不是直接的宏观情感极性. 在 TweetEval 数据集上 (图 4(b)), 实验结果表明对于生气和伤心类别来说, 专家 1 有着明显的偏好, 这说明该专家擅长处理负面情绪, 同时专家 1 处理另外两个正面标签的占比明显较低, 分别对应 0.06、0.09; 专家 3、4 分别擅长处理标签开心和乐观, 这说明这两个专家对应着正面情感专家, 同时这两个专家对于另外两个负面标签的处理率很低. 以上结果表明各个专家在细粒度的情绪方向上的确达到了分化效果. 在 SST-5 数据集上 (图 4(c)), 专家 1 似乎属于情感专家, 其在所有情感类别上均有很高的占比. 其他专家呈现出某种规律: 在两种情感极性的处理率有着明显的差异, 如果擅长处理偏正面的情感, 那负面情感的占用率会很低. 比如对于专家 0 来说, 在标签非常负面到中性的处理率仅有 0.02、0.03 和 0.09, 而在标签正面和非常正面的处理率却达到了 0.31 和 0.49. 这种差异化的激活模式有力地证明了 MoE 架构中的专家通过内部门控机制形成了对特定情感模式的专精. 这种术业有专攻的特性不仅提升了模型的性能, 也为模型的决策过程提供了更强的可解释性: 当模型对某个评论做出情感判



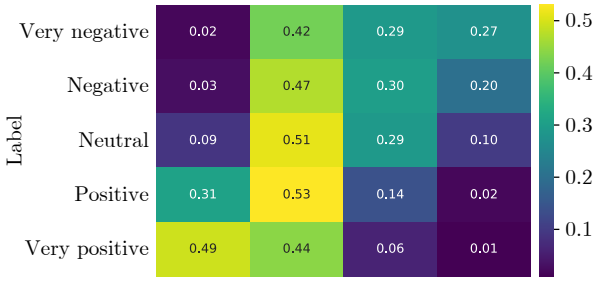
(a) IMDb 激活分析

(a) IMDb activation analysis



(b) TweetEval 激活分析

(b) TweetEval activation analysis



(c) SST-5 激活分析

(c) SST-5 activation analysis

图 4 专家激活模式分析

Fig. 4 Expert activation pattern analysis

断时, 可以追溯到哪些专家在其中发挥了主导作用, 从而理解模型做出该判断的内在逻辑.

3.3.4 专家专业化程度分析

本实验旨在进一步量化和证明 MoE 中的专家确实学习了不同的、非冗余的知识, 而不是简单地复制彼此的功能, 以支持专家专精的论点.

如果 MoE 中的专家是专业的, 那么它们在某种程度上是可区分的. 这种可区分性可以体现在所处理的输入数据上 (输入侧专业化), 也可以体现在对相同或不同输入数据产生的内部表征或输出上 (功能专业化). 输入侧原理: 假设路由网络能够有效

地将不同类型的输入路由到相应的专家, 那么, 被路由到特定专家的输入数据在某种特征空间中聚集在一起. 对这些输入数据进行聚类, 可以检验路由网络是否成功地实现了输入数据的分工. 输出侧原理: 即使输入数据是均匀分布的, 如果专家学习了不同的映射函数, 那么它们对相同输入数据产生的内部表示或输出也应该是不同的. 通过对这些输出进行聚类, 可以直观地看到每个专家关注或擅长的不同方面.

本节在测试集中随机抽取样本, 然后让模型对其进行推理, 推理过程中捕获 MoE 层中各个专家的输入隐层状态以及各个专家处理数据后产生的输出状态, 随后对这些高维向量使用 PCA 进行初步降维 (为了减少噪声, 提高后续可视化的效率, 将向量维度降到 50), 然后利用 t-SNE 把这些向量降到 2 维, 最后把这些向量在 2 维平面进行可视化.

本节选取主实验中性能最好的配置进行实验 (IMDb 和 SST-5 用 4 个专家, TweetEval 选用 6 个专家), 同时只关注最后一层 MoE 的表现, 提供了在三个数据集上的专家专业化程度分析. 对于 IMDb 来说 (图 5(a)), 在输入侧只有专家 3 (紫色) 出现了一定的聚类, 其他专家似乎在输入侧没有形成一定的聚类. 但是相较于输入侧, 输出侧的聚类效果更加明显, 在 2 维平面上形成了与专家数量相同的簇, 出现这一现象是因为专家的确形成了分化. 对于 TweetEval 来说 (图 5(c)、图 5(d)), 其在输入侧表现不佳, 但是在输出侧 (左下角的绿色, 右下角的蓝色) 出现了些许聚类, 其他颜色较为分散, 这可能是由于该数据集中的样本本身比较口语化, 使得各个专家难以捕获不同方面的特征. 对于 SST-5 来说 (图 5(e)、图 5(f)), 其在输入侧只有很小程度的聚类 (紫色), 但是在输出侧的聚类更加明显. 以上结果表明, 不管对于哪个数据集, 输入侧聚类都不是特别明显, 这一现象可能源于两个方面: 一方面, 高维语义空间本身的复杂性使得输入在降维后难以呈现清晰的簇状结构; 另一方面, 当前路由机制对输入样本的划分仍不够精细, 从而限制了潜在簇结构的显现. 但是输出聚类比输入聚类更明显, 这提供了强有力的证据, 表明专家本身正在将特征转换为专门的、可分离的表示形式.

3.3.5　对比消融

本实验旨在说明负载均衡损失的必要性、专家内部设计的门控单元的优势、MoE 模块替换 FFN 层数对性能的影响以及更详细的专家数量分析.

本节依然选择主实验中性能最好的专家配置进行实验. MoE 中的专家模块替换成普通的 FFN, 在三个数据集上的结果如表 4 所示. 结果表明, 不论在哪个数据集上, 采用本文的门控专家, 效果都比传统 FFN 效果好.

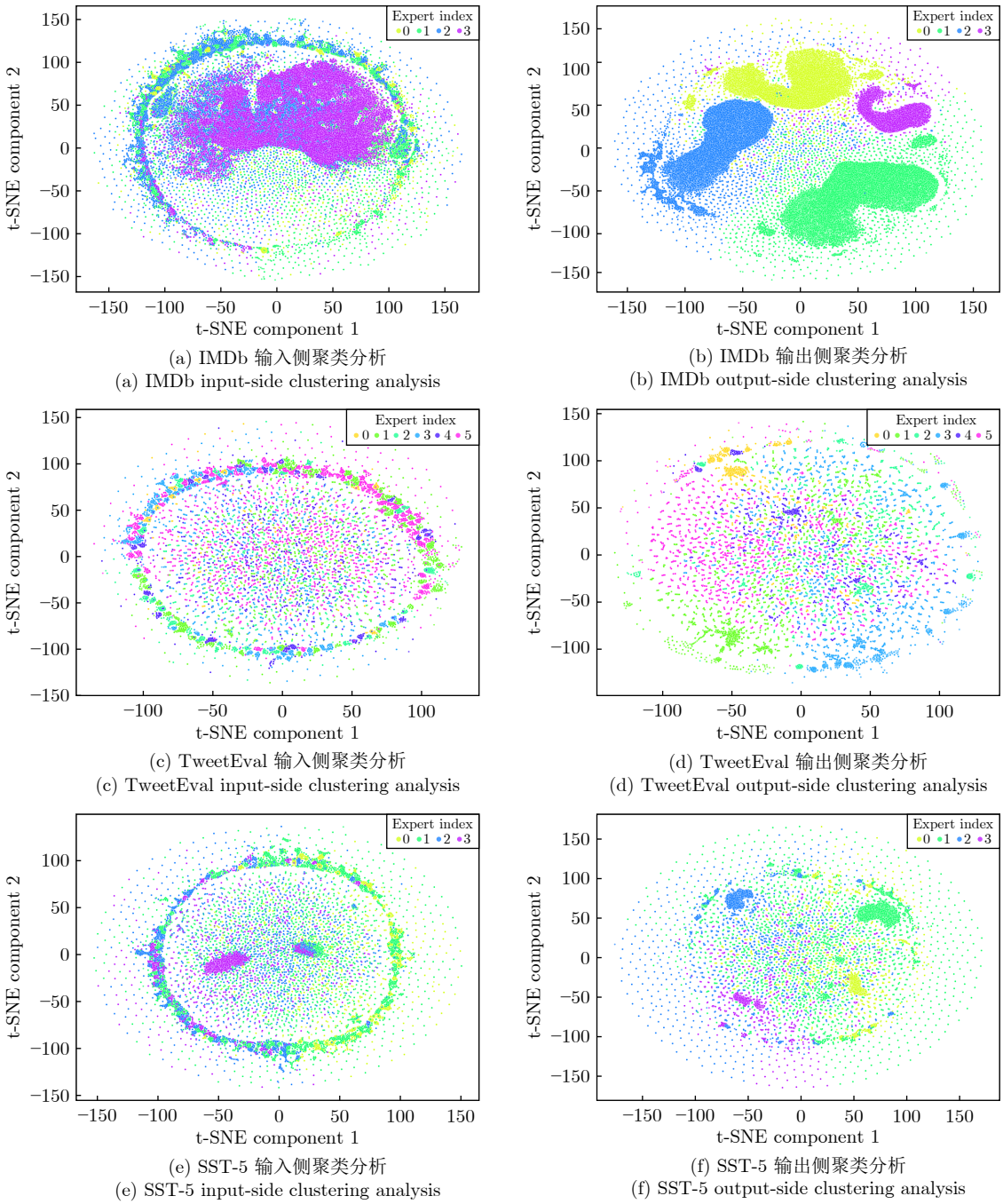
在负载损失实验中, 选用 TweetEval 进行研究 (该数据集上 MoE 为 6 个专家, 专家数量多则更能体现出各个专家的差异性). 把损失函数中的负载损失项去掉, 观察了模型在训练过程中各个专家处理的 token 数. 图 6 实验结果表明当加上负载损失后, 各个专家处理的 token 数基本持平; 当去掉负载项后, 各个专家处理的数量差异巨大. 缺少该机制会导致专家失衡、退化, 模型有效容量下降, 训练表现不佳.

有关专家数量和替换层数对性能的影响见图 7. 在图 7(a) 实验中分别尝试了 1 ~ 8 个专家. 结果表明, 随着专家数量的增加, 模型性能有一定程度的提升. 在 IMDb、SST-5 数据集上专家数量为 4 时达到最优, 而在 TweetEval 上专家数量为 6 时达到最优. 但是随着专家数量的增多, 模型的性能基本不变. 这表明: 当专家数量较少时, 模型缺乏足够的专家从每段文本的不同方面学习关键信息, 因此增加专家数量有助于 MoE 逐步充分学习, 从而提升性能; 当专家数量达到一定规模后, 新增的专家无法进一步带来收益, 模型性能趋于收敛. 在图 7(b) 中尝试了从模型的第 0 层开始替换 (全部替换) 一直到从第 11 层开始替换 (仅仅替换最后一层). 结果表明, 在三个数据集上开始替换的层数对性能的影响基本一致. 具体来说, 若从第 10 层开始替换, 模型性能达到最优. 但随着替换层数的增多模型的性能出现明显的下降, 这表明新增替换层数可能因破坏了预训练模型的中低层语义知识而导致性能下降.

3.3.6　案例分析

前面的主实验、专家激活模式分析和专业化程度分析已经从宏观和中观层面证明了 MoE 架构的有效性和机理. 量化指标只能提供总体趋势, 而具体案例可以从微观层面, 通过具体的、有代表性的例子, 直观地、生动地展示模型是如何工作的, 特别是模型如何处理复杂或微妙的语言现象, 有助于揭示其在处理复杂或边缘情况时的决策过程.

本实验选用的第一个案例是 (图 8(a)) This whole week has been a disaster and I′m so frustrated, but I′m optimistic things will get better from here. 这个例子属于混合情感的情况, 前半段的情绪是负面的, 有生气和沮丧. 而后面的情绪属于正面, 有乐观. 标准的、非 MoE 的架构可能会被其中一种情感带偏, 或者因为情感冲突而感到困惑,



(a) IMDb 输入侧聚类分析
(a) IMDb input-side clustering analysis

(b) IMDb 输出侧聚类分析
(b) IMDb output-side clustering analysis

(c) TweetEval 输入侧聚类分析
(c) TweetEval input-side clustering analysis

(d) TweetEval 输出侧聚类分析
(d) TweetEval output-side clustering analysis

(e) SST-5 输入侧聚类分析
(e) SST-5 input-side clustering analysis

(f) SST-5 输出侧聚类分析
(f) SST-5 output-side clustering analysis

图 5 聚类分析

Fig. 5 Clustering analysis

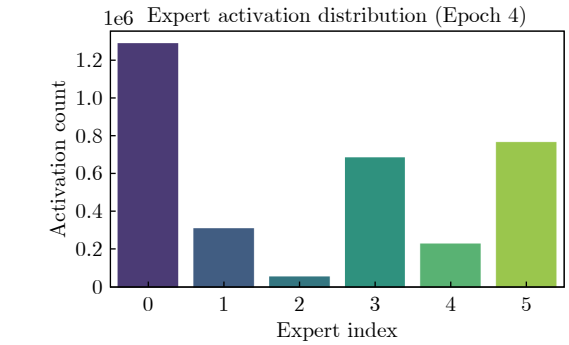
给出一个置信度不高的错误判断. 在所提出的 MoE 架构中: 句子的前半段主要激活专家 0, 其中负面情感词 disaster、frustrated 的激活率为 0.25、0.20, 转折词 but 触发了专家切换, 导致句子的后半段主要被专家 3、4 激活, 其中正面情感词 optimistic, better 对专家 3 的激活率分别为 0.28 和 0.26. 最终模型综合专家意见, 正确预测为积极情感.

第二个案例是 (图 8 (b)) I just love when the app crashes right before a deadline-fantastic. 这个例子属于反讽的情况. 字面上看, 句子中包含了 love 和 fantastic 等积极情感词, 非 MoE 的架构可能会被这些词误导, 从而错误地判断为积极情感. 然而, 人类读者能轻易地识别出其中的反讽意味, 理解其真实的负面情感. 在 MoE 架构中, 专家 2 对 love 和 fantastic 有着较高的激活率 (0.52、0.48), 表明其是一个正向情感专家. 而对于负面词 crashes 和

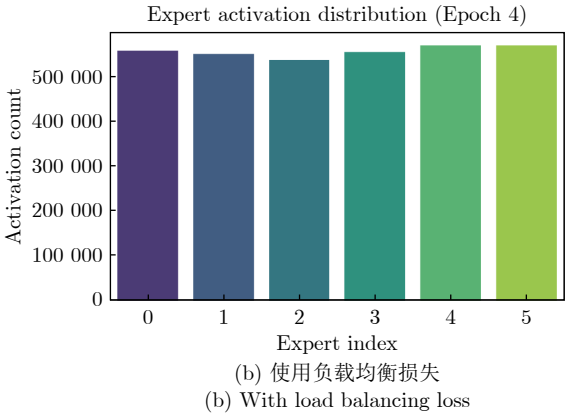
表 4　不同数据集上基线与引入 MoE 后模型的性能对比

Table 4　Performance comparison between baseline and MoE-enhanced models on different datasets

数据集	模型结构	Accuracy	Precision	Recall	F1-Score
IMDb	普通 FFN	0.9551	0.9552	0.9551	0.9551
	门控专家	**0.9565**	**0.9567**	**0.9565**	**0.9565**
TweetEval	普通 FFN	0.8220	0.8222	0.8220	0.8215
	门控专家	**0.8325**	**0.8338**	**0.8325**	**0.8327**
SST-5	普通 FFN	0.5683	0.5677	0.5683	0.5666
	门控专家	**0.5805**	**0.5788**	**0.5805**	**0.5785**



(a) 不使用负载均衡损失
(a) Without load balancing loss



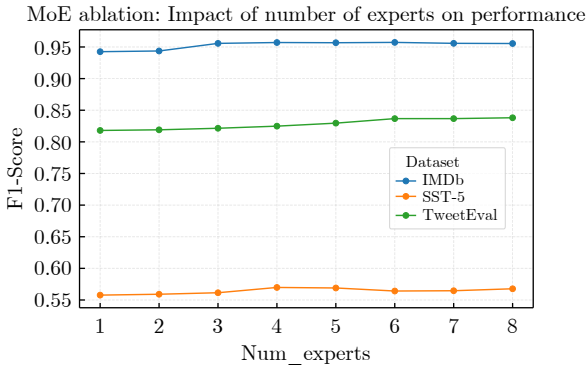
(b) 使用负载均衡损失
(b) With load balancing loss

图 6　负载均衡损失消融实验结果

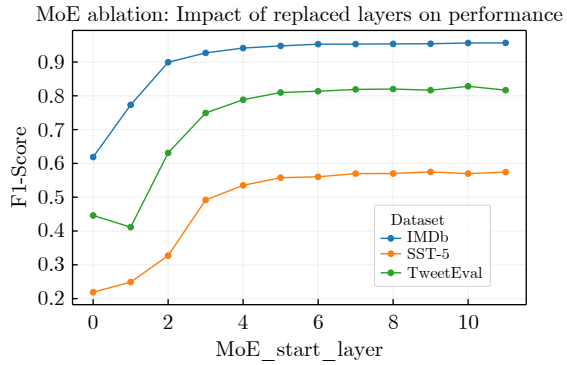
Fig. 6　Ablation experiment results on load balancing loss

deadline 专家 4 有着较高的激活率 (0.42、0.48), 专家之间达到了分工. 最终, 模型综合所有专家的意见, 正确地将这个反讽句子预测为负面情感.

第三个案例是 (图 8(c)) I don′t think it′s not useful to add more logs. 这个例子属于双重否定句, 旨在表达正面情感. 双重否定对于许多模型来说是一个挑战, 它们可能无法正确理解其中的逻辑关系, 从而导致错误的判断. 在 MoE 架构中: 第一个否定词 don′t 主要激活了专家 1 (0.45、0.34), 当



(a) 专家数量消融
(a) Expert count ablation



(b) 替换层数消融
(b) Layer replacement ablation

图 7　消融研究折线图 (层数和专家数量变化)

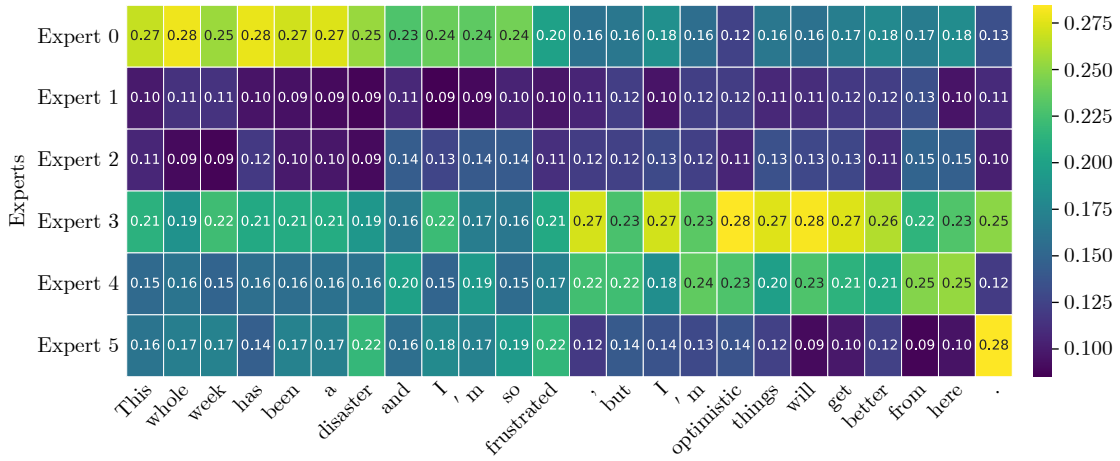
Fig. 7　Ablation study line chart (varying number of layers and number of experts)

遇到第二个否定词 not 时, 模型同样对专家 1 有着较高激活率 (0.53), 这表明该专家专门处理否定词的叠加效应. 而关键字 useful 对除专家 1 外的所有专家有着较为平均的激活率 (0.28、0.16、0.14、0.23、0.17), 整个句子通过不同专家的协作, 准确地理解了双重否定表达的肯定含义, 即添加更多的日志是有用的. 模型综合所有专家的意见, 成功地抵消了双重否定的影响, 做出了正确的判断.

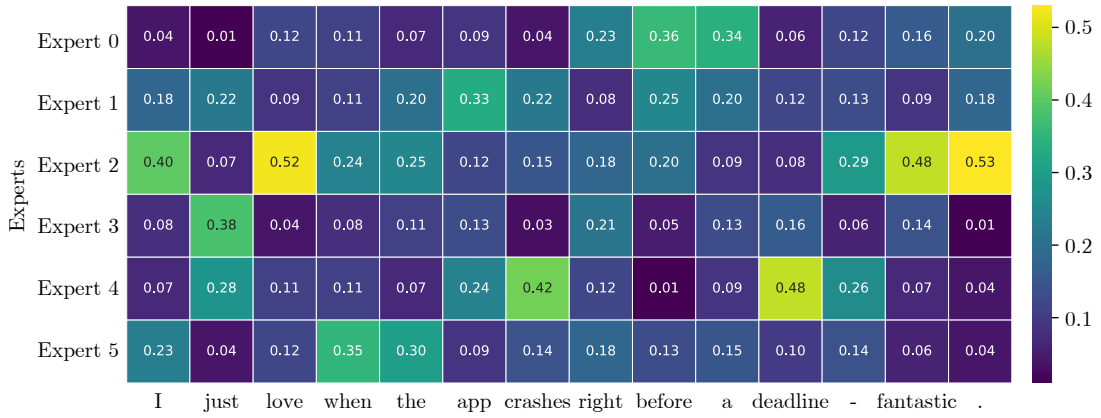
以上案例直观展示了 MoE 可扩展模型如何通过专家分工处理复杂情感表达, 为模型决策提供了可解释的路径. 这种清晰的责任划分表明, 模型并非简单地对整个句子进行粗略判断, 而是能够在 token 级别上识别并分派不同的情感信号, 从而更精准地捕捉文本的复杂语义.

4　结束语

本研究提出一种基于混合专家的可扩展情感分析模型且成功集成到 RoBERTa 中, 并在性能提升、效率分析、专家专精化、可解释性和可扩展性方面进行了深入的实验分析. 通过用自定义 MoE 层替



(a) 案例 1
(a) Case 1



(b) 案例 2
(b) Case 2

(c) 案例 3
(c) Case 3

图 8 案例分析

Fig. 8 Case study

换 RoBERTa 的 FFN 层, 并在每个专家内部引入独特的门控逻辑, 本研究的模型在准确率、精确率、召回率和 F1 分数等关键指标上均超越了基线 RoBERTa 模型和前沿模型. 这表明, 通过引入更细粒度的条件计算, 模型能够学习到更强大、更具区分性的情感表示. 同时实验结果明确证实了专家

专精化的存在, 通过量化专家激活模式和专家重叠程度, 揭示了专家学习不同且非冗余知识的能力. 这种专精化不仅提升了模型性能, 也为模型的可解释性奠定了基础. 案例分析进一步提供直观的证据, 展示了模型如何通过动态激活特定专家来处理复杂的情感输入. 本研究的 MoE-RoBERTa 模型还展现 MoE 架构固有的可扩展性优势. 通过稀疏激活机制, 模型能够在不显著增加计算开销的情况下扩展其总参数量, 展现了其在处理复杂情感任务中的潜力及良好的扩展性, 为未来构建更大规模、更高效、更具可解释性的情感分析系统提供了可能.

未来的工作可以集中在以下几个方面: 1) 更复杂的门控机制. 可以研究更先进的门控单元设计, 例如引入层次化门控、基于上下文的门控机制或自适应门控策略. 2) 多模态情感分析. 将 MoE-RoBERTa 架构扩展到多模态情感分析任务, 利用专家处理文本、图像或音频等不同模态的信息, 以应对日益复杂的用户生成内容. 3) 其他 NLP 任务的推广. 将 MoE-RoBERTa 的成功经验推广到其他 NLP 任务, 如文本分类、命名实体识别、问答系统、文本摘要等, 为构建更通用的 MoE-based NLP 模型奠定基础.

参考文献

1 Bordoloi M, Biswas S K. Sentiment analysis: A survey on design framework, applications and future scopes. *Artificial Intelligence Review*, 2023, **56**(11): 12505−12560

2 Zheng Zhi-Hao, Wu Wen-Bing, Chen Xin, Hu Rong-Xin, Liu Xin, Wang Pu. A traffic sensing and analyzing system using social media data. *Acta Automatica Sinica*, 2018, **44**(4): 656−666
(郑治豪, 吴文兵, 陈鑫, 胡荣鑫, 柳鑫, 王璞. 基于社交媒体大数据的交通感知分析系统. 自动化学报, 2018, **44**(4): 656−666)

3 Wang Hui-Dong, Li Zhao-Dong, Yao Jin-Li, Yu De-Gan. Sentimental propagation model of stock investors based on symmetric triangular fuzzy set. *Acta Automatica Sinica*, 2020, **46**(5): 1031−1043
(王会东, 李兆东, 姚金丽, 余德淦. 基于对称三角模糊集的股票投资者情绪传播模型. 自动化学报, 2020, **46**(5): 1031−1043)

4 He Xin-Run, Li Yi-Xuan, Fu Zhong-Zheng, Wu Dong-Rui, Huang Jian. A study of TSK fuzzy system and domain adaptation method in multi-label affective computing. *Acta Automatica Sinica*, 2025, **51**(7): 1546−1561
(何欣润, 李毅轩, 傅中正, 伍冬睿, 黄剑. 多标签情感计算中的 TSK 模糊系统与域适应方法研究. 自动化学报, 2025, **51**(7): 1546−1561)

5 Baccianella S, Esuli A, Sebastiani F. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In: Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC). Valletta, Malta: European Language Resources Association, 2010. 83−90

6 Li Yu-Qing, Li Xin, Han Xu, Song Dan-Dan, Liao Le-Jian. A bilingual lexicon-based multi-class semantic orientation analysis for microblogs. *Acta Electronica Sinica*, 2016, **44**(9): 2068−2073
(栗雨晴, 礼欣, 韩煦, 宋丹丹, 廖乐健. 基于双语词典的微博多类情感分析方法. 电子学报, 2016, **44**(9): 2068−2073)

7 Zhao Yan-Yan, Qin Bing, Shi Qiu-Hui, Liu Ting. Large-scale sentiment lexicon collection and its application in sentiment classification. *Journal of Chinese Information Processing*, 2017, **31**(2): 187−193
(赵妍妍, 秦兵, 石秋慧, 刘挺. 大规模情感词典的构建及其在情感分类中的应用. 中文信息学报, 2017, **31**(2): 187−193)

8 Yang Shuang, Chen Fen. Analyzing sentiments of micro-blog posts based on support vector machine. *Data Analysis and Knowledge Discovery*, 2017, **1**(2): 73−79
(杨爽, 陈芬. 基于 SVM 多特征融合的微博情感多级分类研究. 数据分析与知识发现, 2017, **1**(2): 73−79)

9 Li J, Rao Y H, Jin F M, Chen H J, Xiang X Y. Multi-label maximum entropy model for social emotion classification over short text. *Neurocomputing*, 2016, **210**: 247−256

10 Alaie A I, Farooq U, Bhat W A, Khurana S S, Singh P. An empirical study on sentimental drug review analysis using lexicon and machine learning-based techniques. *SN Computer Science*, 2024, **5**(1): Article No. 63

11 Wang Ke, Xia Rui. A survey on automatical construction methods of sentiment lexicons. *Acta Automatica Sinica*, 2016, **42**(4): 495−511
(王科, 夏睿. 情感词典自动构建方法综述. 自动化学报, 2016, **42**(4): 495−511)

12 Lai Y N, Zhang L F, Han D H, Zhou R, Wang G R. Fine-grained emotion classification of Chinese microblogs based on graph convolution networks. *World Wide Web*, 2020, **23**(5): 2771−2787

13 Chen L H, Varoquaux G. What is the role of small models in the LLM era: A survey. arXiv preprint arXiv: 2409.06857, 2025.

14 Rezapour M. Emotion detection with Transformers: A comparative study. arXiv preprint arXiv: 2403.15454, 2024.

15 di Palma D, de Bellis A, Servedio G, Anelli V W, Narducci F, di Noia T. LLaMAs have feelings too: Unveiling sentiment and emotion representations in LLaMA models through probing. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vienna, Austria: Association for Computational Linguistics, 2025: 6124−6142

16 Chen K Z, Wang S, Ben H X, Tang S G, Hao Y B. Mixture of multimodal adapters for sentiment analysis. In: Proceedings of the Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). Albuquerque, New Mexico: Association for Computational Linguistics, 2025. 1822−1833

17 Lo K M, Huang Z Y, Qiu Z H, Wang Z L, Fu J. A closer look into mixture-of-experts in large language models. *Findings of the Association for Computational Linguistics: NAACL 2025*, 2025. 4427−4447

18 Mu S Y, Lin S. A comprehensive survey of mixture-of-experts: Algorithms, theory, and applications. arXiv preprint arXiv: 2503.07137, 2025.

19 Nnamdi J, Dimitri V, Amar S. Improving deep learning performance with mixture of experts and sparse activation. *Preprints 2025*, DOI: 10.20944/preprints202503.0611.v1

20 Nguyen H, Ho N, Rinaldo A. On least square estimation in softmax gating mixture of experts. In: Proceedings of the 41st International Conference on Machine Learning (ICML). Vienna, Austria: PMLR, 2024. 37707−37735

21 Wang K, Shen W Z, Yang Y Y, Quan X J, Wang R. Relational graph attention network for aspect-based sentiment analysis. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Virtual Event: Association for Computational Linguistics, 2020. 3229−3238

22 Talaat A S. Sentiment analysis classification system using hybrid BERT models. *Journal of Big Data*, 2023, **10**(1): Article No. 110

23 Krishnamoorthy A, Sundhar K A, Naveen K V, Karthik V. Analyzing sentiments: A comprehensive study of Roberta-based sentiment analysis on twitters. In: Proceedings of the 4th International Conference on Advancement in Electronics & Commu-

nication Engineering (AECE). Ghaziabad, India: IEEE, 2024. 626−630

24 Cai W L, Jiang J Y, Wang F, Tang J, Kim S, Huang J Y. A survey on mixture of experts in large language models. *IEEE Transactions on Knowledge and Data Engineering*, 2025, **37**(7): 3896−3915

25 Shazeer N, Mirhoseini A, Maziarz K, Davis A, Le Q, Hinton G, et al. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In: Proceedings of the 5th International Conference on Learning Representations (ICLR). Toulon, France: OpenReview.net, 2017.

26 Fedus W, Zoph B, Shazeer N. Switch Transformers: Scaling to trillion parameter models with simple and efficient sparsity. *The Journal of Machine Learning Research*, 2022, **23**(1): Article No. 120

27 Du N, Huang Y P, Dai A M, Tong S, Lepikhin D, Xu Y Z, et al. GLaM: Efficient scaling of language models with mixture-of-experts. In: Proceedings of the 39th International Conference on Machine Learning (ICML). Baltimore, USA: PMLR, 2022. 5547−5569

28 Zhu T, Qu X Y, Dong D Z, Ruan J C, Tong J Q, He C H, et al. LLaMA-MoE: Building mixture-of-experts from LLaMA with continual pre-training. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Miami, USA: Association for Computational Linguistics, 2024. 15913−15923

29 Tairin S, Mahmud S, Shen H Y, Iyer A. eMoE: Task-aware memory efficient mixture-of-experts-based (MoE) model inference. arXiv preprint arXiv: 2503.06823, 2025.

30 Liu Y H, Ott M, Goyal N, Du J F, Joshi M, Chen D Q, et al. RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv: 1907.11692, 2019.

陈 千 山西大学计算机与信息技术学院副教授. 主要研究方向为自然语言处理, 情感计算.
E-mail: chenqian@sxu.edu.cn
(**CHEN Qian** Associate professor at the School of Computer and Information Technology, Shanxi University. His research interests include natural language processing and affective computing.)

胡梦强 山西大学计算机与信息技术学院硕士研究生. 主要研究方向为自然语言处理, 情感计算. 本文通信作者.
E-mail: leep87233@gmail.com
(**HU Meng-Qiang** Master student at the School of Computer and Information Technology, Shanxi University. His research interests include natural language processing and affective computing. Corresponding author of this paper.)

郭 鑫 山西大学计算机与信息技术学院副教授. 主要研究方向为自然语言处理, 情感计算.
E-mail: guoxinjsj@sxu.edu.cn
(**GUO Xin** Associate professor at the School of Computer and Information Technology, Shanxi University. Her research interests include natural language processing and affective computing.)

王素格 山西大学计算机与信息技术学院教授. 主要研究方向为自然语言处理, 机器学习.
E-mail: wsg@sxu.edu.cn
(**WANG Su-Ge** Professor at the School of Computer and Information Technology, Shanxi University. Her research interests include natural language processing and machine learning.)