SentiFuse：用于稳健情感提取的深度多模型融合框架

Hieu Minh Duong [1]，Rupa Ghosh [1]，Cong HoanNguyen [1]，Eugene Levin [2]，ToddGary [2]，以及 Long Nguyen [1]

[1]美国肯塔基州路易斯维尔市路易斯维尔大学
{hieu.duong,rupa.ghosh,conghoan.nguyen,l.nguyen}@louisville.edu[2]美国田纳西州纳什维尔市梅哈里医学院 {tpgary,elevin}@mmc.edu

arXiv:2602.01447v1 [cs.CL] 1 Feb 2026

摘要

情感分析模型各有所长，但现有方法缺乏有效整合这些模型的统一框架。我们提出了SentiFuse，这是一种灵活且与模型无关的框架，通过标准化层和多种融合策略整合异构情感模型。该方法支持决策级融合、特征级融合和自适应融合，从而实现对不同模型的系统性组合。我们在三个大型社交媒体数据集 Crowdflower、GoEmotions 和 Sentiment140 上进行了实验。结果表明，SentiFuse 的表现始终优于单个模型和朴素集成。特征级融合的整体效果最强，相较于最佳单体模型和简单平均，F1分数最高提升了 4 个百分点；而自适应融合则增强了模型在否定、混合情绪和复杂情感表达等挑战性案例中的鲁棒性。这些结果表明，系统性地利用模型间的互补性，能够在不同数据集和文本类型上实现更准确、可靠的情感分析。

K关键词： 情感分析、模型融合、文本分类、自然语言处理 .

1 引言

大语言模型（LLMs）和深度学习架构通过捕捉传统方法难以企及的语义依赖关系和语境细微差别，改变了情感分析。VADER 等基于规则的系统 [15] 具有可解释性和效率优势，但在面对讽刺、领域特定语言或复杂修饰语时往往表现脆弱。统计模型计算开销较低，却无法充分捕捉序列依赖关系和上下文关系 [12]。相比之下，深度学习模型擅长对语义进行建模，但需要大量带标签的数据集，并存在过拟合风险 [24]。这些互补的优势与劣势在社交媒体中尤为明显，因为情感表达高度依赖上下文，且常包含反讽、混合情绪或语用线索，这些因素都对单一模型方法构成挑战 [22, 25]。例如，基于词典的模型可能会将句子“太棒了，又一趟航班延误——这正是我今天需要的！”误判为正面情感，因为其中含有“太棒了”和“需要”等词；而上下文敏感模型则能够识别其中的讽刺意图。此类案例既凸显了单一情感分析方法的局限性，也展现了系统整合异构模型、利用其互补视角所能带来的潜在收益。

尽管集成学习前景广阔，现有的情感分析融合方法通常仍较为简单且缺乏系统性。多数投票、无权平均和置信度加权方案都假定基模型同质、输出兼容且决策边界一致 [16, 21]。

*通讯作者：hieu.duong@louisville.edu

被接收至 第14届计算数据与社会网络国际会议（CSoNet 2025） .

这些朴素方法忽视了异构模型集成中的根本挑战，例如输出格式（概率、逻辑值、离散标签）不一致、置信度校准存在差异，以及缺乏能够适应不同语言复杂性的自适应机制。为弥补这些不足，我们提出 SentiFuse，这是一种模型无关框架，通过标准化输出处理和多种融合策略集成异构情感模型。我们的工作聚焦于三个研究问题：系统化融合是否优于朴素组合方法？（RQ1）、融合效果如何随文本特征而变化？（研究问题2）、该框架能否推广到不同的模型组合？（RQ3）。为此，我们的贡献主要有三点。

- 首先，我们引入一个标准化层，将不同的输出统一为概率分布，从而无需进行架构修改即可实现集成。

- 其次，我们设计了三种互补的融合策略——使用学习得到的权重进行决策级融合、利用元分类进行特征级融合，以及由自动提取的文本特征引导的自适应融合。

- 最后，我们围绕三个研究问题开展了全面评估，结果表明，在多样且复杂的情感场景中，SentiFuse 的性能优于单独使用的模型和传统集成模型。

2 相关工作

早期的情感分析依赖词频-逆文档频率向量化等统计方法，并结合传统分类器 [4, 6, 9, 26], ；这些方法效率较高，但忽略了序列依赖关系和语义。VADER 等基于规则的系统 [15] 以及 SentiWordNet [1] 借助语言学启发式规则提高了可解释性，但在上下文敏感性和领域适应方面表现不佳 [5, 11, 23, 29]。深度学习模型 [2, 27, 31, 33] 则解决了这些问题：双向长短期记忆网络等循环架构能够捕捉长距离依赖关系 [32], ，而注意力机制则突出情感承载片段 [34]。以 BERT [8] 和 RoBERTa [18], 为代表的 Transformer 模型通过双向上下文编码取得了最先进的结果，但代价是更高的数据需求，且对领域偏移较为敏感。

除单个模型外，集成方法也得到了广泛应用。装袋法 [3] 和投票法 [10] 等传统方法提升了鲁棒性，但未能充分利用异构系统。较新的元集成方法 [17] 和自适应融合策略 [13]能够动态地为模型输出或模态分配权重，从而取得更好的效果，但通常需要大量计算资源或配对多模态数据。与此同时，大语言模型（大语言模型）的兴起改变了情感分析研究的方向：研究报告显示，与微调后的 Transformer模型 [19]相比，ChatGPT 在零样本和少样本场景 [30] 中取得了具有竞争力甚至更优的表现。跨语言集成模型通过结合翻译与 Transformer 集成模型 [20], 也展现出出色的效果，但误差传播仍是一项挑战。近期综述强调，许多融合方法仍依赖静态加权，且缺乏针对更简单基线的系统评估 [28], ，这凸显了对模型无关的自适应框架的需求。

3 方法论

本研究提出了一种创新的多模型框架 SentiFuse，旨在整合多个异构情感分析模型。我们提出的情感分析框架由四个相互关联的组件构成：（1）多个异构情感分析模型；（2）标准化层；（3）融合策略选择器；以及（4）情感分类。该结构化方法能够灵活整合并协调组合不同模型类型产生的情感预测，涵盖从简单的基于词典的方法到复杂的语言模型。

3.1 多模型情感分析。

第一个组件由一组多样化的情感模型组成，表示如下：

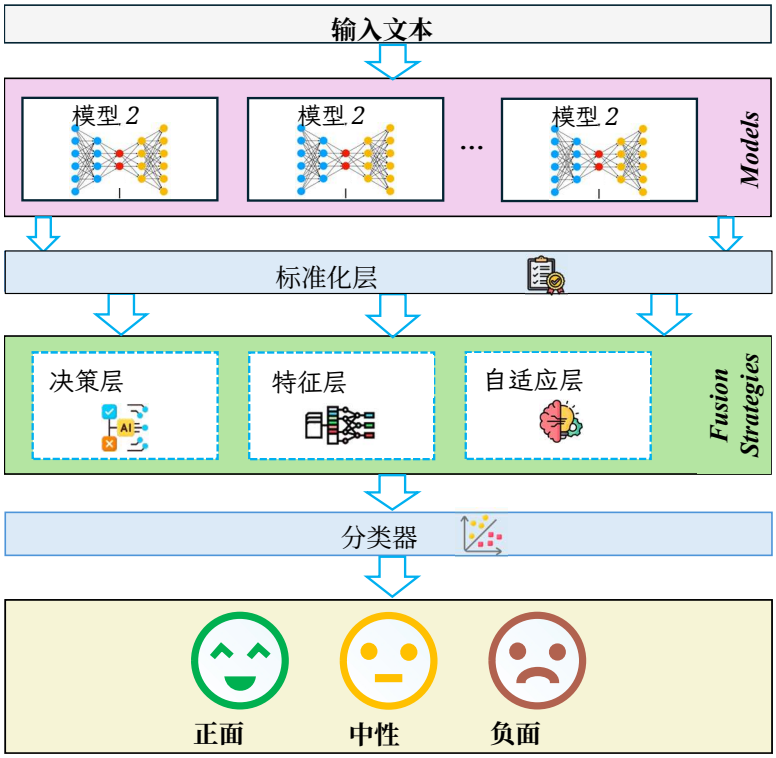

图1：所提出的 SentiFuse 框架的整体架构。

$$M = M_1, M_2, ..., M_n. \tag{1}$$

每个模型 M_i 处理输入文本并生成输出 $O_i = M_i(x)$。我们的实现具体包括：

- **基于词典的模型：**通过汇总精选词典中的词级极性来估计情感，通常会使用启发式规则处理强度、位置强调和效价转换词（例如否定）。
- **基于模式的模型：**根据预定义的情感承载模式（例如习语、表情符号序列、依存模板）的出现情况及其强度推断情感，并根据模式的频率和可靠性为其赋权。

- **机器学习模型：**使用监督分类器（例如逻辑回归、支持向量机），学习从人工设计特征（例如 n元语法、TF–IDF、句法线索、词典特征）到情感标签的映射。

- **编码模型：**使用深度上下文编码器（例如 BERT、RoBERTa）获取序列表示；经过预训练和微调后，将池化向量（例如 [CLS] 标记）传递给特定任务的分类器。

我们将基于人工设计特征训练的经典分类器与学习上下文表示的神经编码模型区分开来；对于分类，编码器会搭配特定任务的预测头并进行微调。

3.2 标准化层。

为便于无缝整合异构模型输出，我们实现了一个标准化函数 S，将不同模型的输出转换为情感类别上的统一概率分布，如下所示：

$$S(O_i) = \begin{cases} \{p_{\text{pos}}, p_{\text{neg}}\}, & \text{if output is probability} \\ \left\{\frac{1+s}{2}, \frac{1-s}{2}\right\}, & \text{if output is score } s \in [-1,1] \\ \{\sigma(v_{\text{pos}}), \sigma(v_{\text{neg}})\}, & \text{if output is logits} \end{cases}. \tag{2}$$

此外，我们为每种模型类型定义了相应的特征提取函数 $\phi_i(O_i)$，如下所示：

$$\phi_i\left(O_i\right)=\left[f_1\left(O_i\right), f_2\left(O_i\right), \ldots, f_k\left(O_i\right)\right]^{\top}, \tag{3}$$

其中，每个 f_i 表示与模型类型相关的特征提取操作。

3.3 融合策略。

我们将三种不同的融合方法形式化如下：

- **决策级融合**：使用加权平均组合各模型的标准化概率输出，定义如下：

$$F_{\mathrm{d}}\left(S\left(O_1\right), \ldots, S\left(O_n\right)\right)=\frac{\sum_{i=1}^{n} w_i \cdot S\left(O_i\right)}{\sum_{i=1}^{n} w_i}, \tag{4}$$

其中，模型特定权重为 $w_i \in [0,1]$。

- **特征级融合**：将多个模型提取的特征聚合为一个用于分类的统一向量，表示如下：

$$F_{\mathrm{f}}\left(S\left(O_1\right), \ldots, S\left(O_n\right)\right)=g\left(\left[\phi_1\left(O_1\right) \oplus \ldots \oplus \phi_n\left(O_n\right)\right]\right), \tag{5}$$

其中，$\oplus$ 表示拼接，ϕ_i 是模型 i的特征映射，而 g 是一个经过训练的元分类器

- **自适应融合：** 根据定义的文本特征动态调整各模型的权重：

$$F_{\mathrm{a}}(x, S(O_1), \ldots, S(O_n))=\frac{\sum_{i=1}^{n} w_i(x) \cdot S(O_i)}{\sum_{i=1}^{n} w_i(x)}, \tag{6}$$

其中，$w_i(x)$ 是根据文本特征 $\psi(x)$ 确定的自适应权重（例如，否定词出现情况、文本长度、情感复杂度）。

3.4 情感分类。

融合后的输出通过分类函数映射为情感概率，定义如下：

$$C\left(F\left(S\left(O_1\right), \ldots, S\left(O_n\right)\right)\right)=\left\{p_{c_1}, p_{c_2}, \ldots, p_{c_k}\right\}, \tag{7}$$

其中，p_{c_i} 是情感类别 c_i的预测概率。最终的情感标签 $L(x)$由置信度阈值 δ确定。例如，如果数据集有以下三种标签：

$$L(x)=\begin{cases}\text{positive}, & p_{\text{pos}}>p_{\text{neg}}+\delta \\ \text{negative}, & p_{\text{neg}}>p_{\text{pos}}+\delta \\ \text{neutral}, & \text{otherwise}\end{cases}. \tag{8}$$

这种广义公式使我们的框架能够无缝处理多类别情感任务，并在多种情感分析场景中提供更大的灵活性。

3.5 训练集与适配。

融合权重通过带标签的数据训练。决策级融合在验证集上调整权重，而特征级融合则采用带 L2 正则化的逻辑回归。自适应融合根据文本类型调整权重：对于包含否定或混合情绪的文本，赋予 Transformer 模型更高权重；对于短文本，则赋予词典更高权重。各模型初始权重相同，再通过简单规则进行调整，使系统保持高效且易于扩展。

4 实验与评估

4.1 实验设置

4.1.1 数据集。

我们在三个情感数据集上评估 SentiFuse：

• **Crowdflower 美国航空推特数据集** （14.6k 条推文）：一个标签均衡（正面、负面、中性）的基准数据集，聚焦于航空客户体验。

• **GoEmotions** （211k 条 Reddit帖子） [7]：谷歌提供的综合情绪数据集，包含 28 种细粒度情绪，并归类为正面、负面和中性情感类别。

• **Sentiment140** （1.6M 条推文） [14]：一个用于情感分类的大规模推文数据集，标签为负面、中性、正面。

所有数据集均采用一致的预处理流程，包括文本规范化和情感标签标准化。我们采用 80-10-10 的分层划分，将数据分为训练集、验证集和测试集。

4.1.2 基线模型。

为评估我们的框架，我们有意选用一组多样化的情感分析模型，以代表不同的方法论范式：

• **经典机器学习模型**。我们采用 TF–IDF 向量化，并结合支持向量机和 XGBoost 等线性分类器。这些模型依赖词袋式表示，能够有效捕捉词汇模式，但会忽略深层上下文语义。

• **深度神经网络模型**。为代表最先进的上下文嵌入模型，我们纳入了 BERT、RoBERTa 和 DistilBERT。实验中使用的版本均为 Hugging Face Transformers 提供的不区分大小写版本，并在英文维基百科和 BookCorpus 上进行了预训练。我们未进行额外预训练，而是直接在情感数据集上对这些模型进行微调。这些 Transformer 模型能够编码丰富的语义和句法信息，因此在各项基准测试中表现出色，但计算成本较高。

为进行公平比较，我们还评估了几种标准集成规则：简单平均、置信度加权平均、多数投票、中位数平均和最大置信度选择。

4.2 研究问题1

第一个研究问题“系统化融合是否优于朴素组合方法？”评估结构化融合策略相较于朴素集成规则和单个模型是否能够带来可测量的收益。本分析所采用的固定模型池由**VADER**、**DistilBERT**和**TF–IDF** 分类器组成，分别代表基于词典的、神经网络和统计范式。我们比较三类方法：(i) 表现最佳的单个模型（通常为 DistilBERT）；(ii) 朴素集成，包括简单平均、置信度加权平均、多数投票、中位数平均和最大置信度选择；以及 (iii) 结构化融合方法，即决策级融合、特征级融合和自适应融合。表1报告了各策略在不同数据集上的准确率、精确率、召回率和 F1分数。

在**Crowdflower**上，获得了最高准确率（90.71）和较高的精确率（73.99）。多数投票取得了最佳 F1值（69.94），略高于特征级融合（69.51）以及决策级融合/简单平均（69.53）。在**GoEmo-tions**上，特征融合取得了最高准确率（78.49）。值得注意的是，中位数平均取得了最佳 F1值（69.19），超过了决策级融合/简单平均（67.66）和特征级融合（67.23）。在大规模的**Sentiment140**上，特征融合在准确率（77.85）和 F1值（77.89）两项指标上均表现最佳，优于最佳单模型（76.11/76.43）。在各数据集上，特征级融合总体而言是最可靠的优胜方法（在三个数据集上均取得最佳准确率；在 Sentiment140 上取得最佳 F1值）。鉴于其规模庞大（160万样本）且具有平衡的类别分布，我们还在图2 中专门展示了 Sentiment140 数据集的 ROC 曲线和 PR 曲线。

特征级融合表现最出色，ROC-AUC 为 0.856，PR 曲线下面积为 0.857，优于最佳单体模型（AUC 为 0.837，PR 曲线下面积为 0.837）。决策融合和简单平均紧随其后（0.834–0.830），而多数投票和最大置信度方法则表现较差

表 1：各数据集上融合策略的性能（百分比）。

策略	Crowdflower				GoEmotions				Sentiment140			
	召回率	Prec	Acc	F1	召回率	Prec	Acc	F1	召回率	Prec	Acc	F1
最佳单模型	77.80	58.79	87.60	66.97	53.59	**77.29**	77.32	63.29	77.28	75.60	76.11	76.43
简单平均	**80.34**	61.29	88.63	69.53	**71.61**	64.12	75.01	67.66	69.40	76.99	74.27	73.00
置信度加权	78.44	61.73	88.66	69.09	68.93	63.51	74.21	66.11	65.51	77.78	73.33	71.12
多数投票	70.82	69.07	90.16	**69.94**	59.86	74.43	77.85	66.36	53.92	**79.95**	70.13	64.40
中位数平均	79.49	60.65	88.35	68.80	71.26	67.24	76.84	**69.19**	71.94	75.80	74.42	73.82
最大置信度	77.59	61.68	88.59	68.73	67.07	62.91	73.55	64.92	62.40	77.67	72.16	69.20
决策融合（我们的方法）	**80.34**	61.29	88.63	69.53	**71.61**	64.12	75.01	67.66	69.40	76.99	74.27	73.00
特征融合（我们的方法）	65.54	**73.99**	**90.71**	69.51	60.46	75.70	**78.49**	67.23	**77.85**	77.93	**77.85**	**77.89**
自适应融合（我们的方法）	**80.34**	60.22	88.25	68.84	70.95	63.43	74.47	66.98	68.27	77.25	74.02	72.48

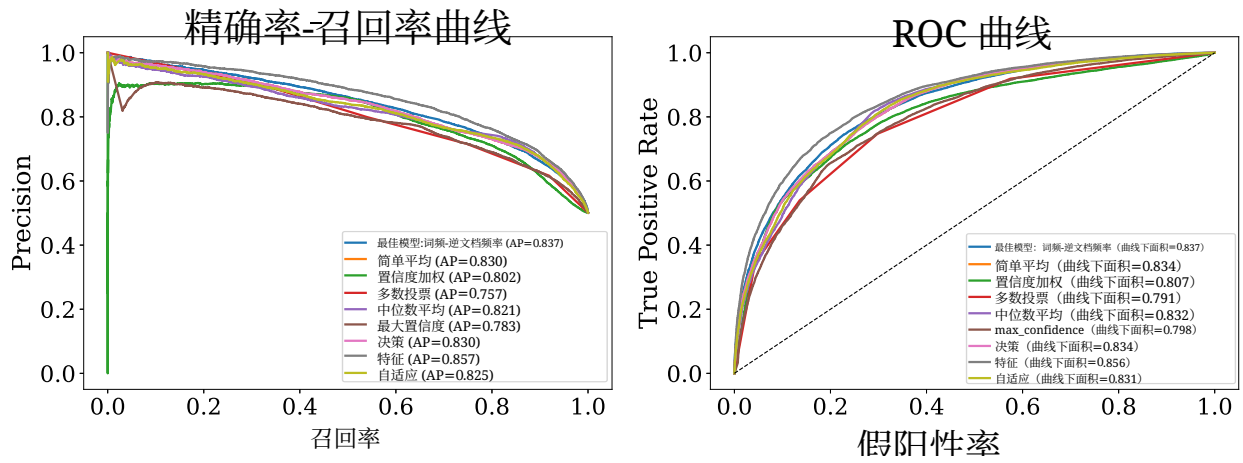

图 2：Sentiment140 数据集上的 PR 曲线和 ROC 曲线。

（AUC 低于 0.800，PR 曲线下面积最低为 0.757）。这些结果证实，结构化融合比朴素集成规则具有更强的判别能力和更可靠的精确率–召回率权衡。

4.3 研究问题 2

第二个研究问题“融合效果如何随文本特征而变化？”考察不同融合策略是否在特定类型的文本上表现更好。鉴于模型各有互补优势，例如，词典方法擅长处理短文本，而 Transformer 模型能够捕捉复杂语义，我们假设，与简单规则或单个模型相比，结构化融合能够更有效地适应文本条件。随后，我们比较了每个类别中的最佳单体模型、决策融合、自适应融合和特征融合。图 3 报告了三个数据集上的准确率。

在 **Crowdflower** 上，特征融合在几乎所有类别中均持续优于其他方法。它在否定上达到约 0.95 的准确率，在长文本上接近 1.0，表明模型组合能够捕捉细微的极性转变，并受益于更丰富的上下文。即便在基于词典的方法通常表现优异的短文本上，特征融合也达到 0.82，超过决策（0.75）、自适应（0.76）和最佳单体模型（0.75）。这表明特征级集成可以通过汇集互补特征来弥补稀疏信号。对于 **GoEmotions**，特征融合的优势再次显现，尤其是在否定（0.77）和复杂文本（0.73）等具有挑战性的类别中。短文本同样更有利于特征融合（0.79，而决策为 0.77，最佳单模型为 0.73）。有趣的是，在混合情绪上，各策略之间的差距缩小，表明所有模型都认为这一现象本身难以处理，而融合仅能部分缓解这一挑战。在长文本上，所有方法都收敛到接近 1.0，反映出在上下文充足时，单个模型和集成模型都能稳健表现。在大规模 **Sentiment140** 上，特征融合在各类别中保持稳定优势，但优势幅度小于其他数据集。它在短文本上得分为 0.74，而最佳单模型为 0.73；在复杂文本上为 0.69，而决策和

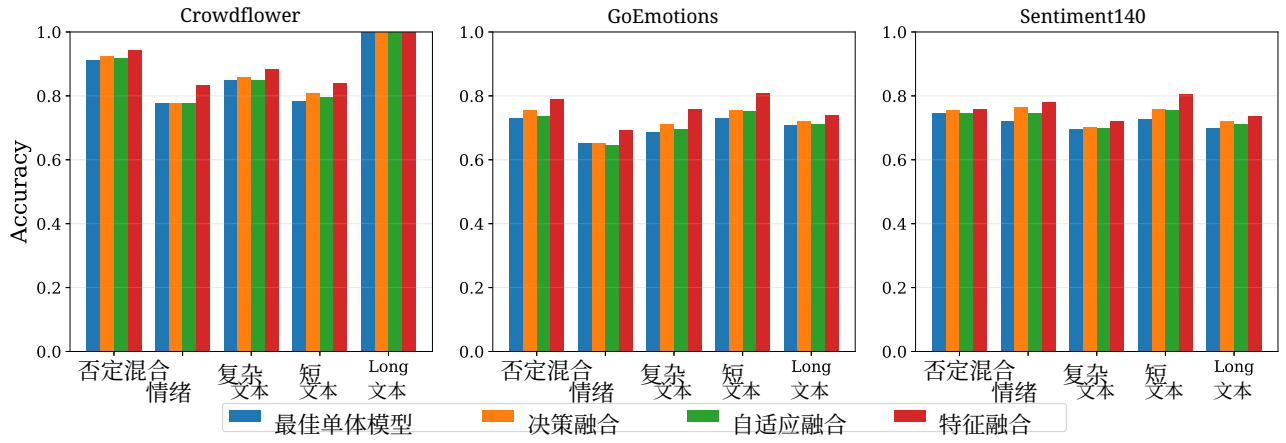

图 3：所提出框架与最佳单体模型在各数据集上的准确率。

自适应融合为 0.68。这表明，在超大规模数据集上，强大的单个模型已经捕捉了大部分可用信号，而融合策略带来的是增量但可靠的改进。

总体而言，研究问题2证实，融合的有效性因文本特征而异。对于否定、复杂情感和短文本，特征融合最为有效，因为单个模型在这些情况下表现不佳。决策融合和自适应融合表现稳定，通常接近特征融合，但并未始终超越它。在长文本上，所有方法的表现趋于一致，这表明当输入有限或含义模糊时，融合最能发挥价值。

4.4 研究问题3

第三个研究问题：“该框架能否推广到不同的模型池？”，旨在考察所提出的框架应用于不同异构模型组合时是否仍然有效。在三个数据集上，我们的结果表明，系统化融合始终能达到或优于朴素集成，见表2 和表3。

TextBlob + RoBERTa +SVM组合在各数据集上的融合策略性能（百分比）。

策略	Crowdflower				GoEmotions				Sentiment140			
	召回率	Prec	Acc	F1	召回率	Prec	Acc	F1	召回率	Prec	Acc	F1
简单平均	58.33	73.25	78.11	64.68	62.92	71.06	75.90	**66.70**	52.02	75.58	64.17	56.33
置信度加权	59.33	72.69	78.11	64.92	62.05	71.53	75.94	66.39	49.61	75.61	63.65	54.98
多数投票	**62.83**	72.65	**80.43**	**66.91**	59.84	**72.35**	75.97	65.53	49.34	76.06	63.60	54.95
中位数平均	60.67	73.09	79.18	66.30	61.52	71.71	75.98	66.25	**52.14**	75.55	**64.37**	**56.40**
最大置信度	60.17	72.53	78.93	65.68	59.73	71.95	75.60	65.09	45.86	**76.24**	62.61	52.07
决策融合（我们的方法）	57.00	74.18	77.89	64.27	**62.65**	71.37	75.75	66.26	51.63	75.11	63.56	56.12
特征融合（我们的方法）	61.67	73.67	79.80	66.61	61.92	71.56	**76.09**	66.18	51.49	75.68	63.63	55.93
自适应融合（我们的方法）	55.50	**75.83**	77.39	64.13	62.06	71.99	75.96	66.16	50.39	75.46	63.58	55.79

在**TextBlob + RoBERTa + SVM (Combo 2)**组合中，朴素集成在不同数据集上的表现不一。多数投票在 Crowdflower 上取得最高准确率（80.4%），在 F1 值上也表现最佳（66.9%）；而中位数平均在 GoEmotions 上表现最佳（F1 值为 66.3%）。这些结果表明，朴素集成对数据集条件较为敏感，难以稳定地泛化。相比之下，SentiFuse 策略的表现更为稳定。特征融合在各数据集上均取得了有竞争力的结果（在 Crowdflower 上的准确率为 79.8%，F1 值为 66.6%），而自适应融合在 Crowdflower 上达到最高精确率（75.8%），表明其在偏斜分布下具有鲁棒性。决策融合表现依然可靠，与简单平均的结果十分接近，但方差更低。在**AFINN +BERT + XGBoost (Combo 3)**组合中，结构化融合展现出更明显的优势。特征融合整体表现最为均衡，包括在 Crowdflower 上达到 90.4% 的准确率，以及在 GoEmotions 上达到 75.97% 的精确率和 78.59% 的准确率，优于朴素集成和单个模型。自适应

表3：结合 AFINN + BERT + XG- Boost 的融合策略在各数据集上的表现（百分比）。

策略	Crowdflower				GoEmotions				Sentiment140			
	召回率	Prec	Acc	F1	召回率	Prec	Acc	F1	召回率	Prec	Acc	F1
简单平均	75.66	61.94	88.16	68.20	71.08	64.37	75.20	67.29	66.78	76.07	73.63	71.59
置信度加权	75.16	62.19	88.14	68.20	70.09	64.99	75.18	67.23	63.23	76.40	72.46	70.07
多数投票	71.00	67.62	89.91	69.02	62.59	72.98	77.83	66.59	56.29	**78.17**	70.22	65.34
中位数平均	74.00	62.38	87.89	67.80	70.83	66.18	76.36	67.27	69.85	75.71	74.71	72.70
最大置信度	73.66	62.47	88.20	67.90	69.05	64.91	75.14	66.60	60.55	76.89	72.06	68.38
决策融合（我们的方法）	75.66	61.94	88.16	68.20	71.07	64.37	75.20	67.29	66.78	76.07	73.63	71.59
特征融合（我们的方法）	63.83	**72.69**	**90.39**	**69.08**	61.10	**75.97**	**78.59**	67.39	**77.07**	76.89	**76.89**	**76.90**
自适应融合（我们的方法）	**79.16**	60.62	88.27	68.51	**72.52**	63.67	75.18	**67.82**	66.44	76.89	73.49	71.36

融合侧重于召回率，在 Crowdflower 上达到 79.2%；决策融合则再次表现稳定，与简单平均的结果相符。在 Sentiment140 上，特征融合取得了最高的整体表现（准确率和 F1 值均为 76.9%），这表明即使在大规模场景下，结构化集成仍能带来优势。

总体而言，RQ3 表明 SentiFuse 能够在异构模型融合中实现有效的泛化。尽管朴素集成有时在个别情况下表现良好，但其结果在不同数据集之间存在较大差异。相比之下，结构化方法能够持续提升可靠性：特征融合在细粒度场景中表现突出，自适应融合增强了召回率和鲁棒性，而决策融合则提供了稳定的性能。这些结果凸显出，该框架并不局限于特定的模型集合，而是提供了一种用于整合多样化情感分类器的通用架构。

5 结论

本文介绍了 SentiFuse，这是一种灵活且模型无关的情感分析框架，通过统一的标准化与融合流水线整合多种模型。该框架支持决策级、特征级和自适应融合策略，能够提升其在多个数据集以及否定、混合情绪、简短或复杂表达等具有挑战性的文本现象上的表现。实验表明，没有哪一种策略在所有情况下都占据优势：特征级融合总体上带来显著提升，而自适应和决策级方法则能在异质上下文中增强鲁棒性。特征级融合在否定和混合情绪等类别上尤其出色，因为词典与神经嵌入中的互补线索能够相互印证。自适应融合则通过动态调整模型权重，往往能更好地泛化到不同类别。这些发现凸显了本研究在方法论上的贡献：提供了一种可应用于不同模型族的通用融合架构；在分析层面揭示了特定融合策略在何种文本类型上奏效及其原因；在实践层面则表明，这些策略能够有效泛化到不同的模型池。这些贡献共同为情感分类提供了更可靠的基础，并指明了鲁棒性与互补性至关重要的更广泛应用方向。尽管 SentiFuse 需要运行多个模型，因此会增加一些推理成本，但它仍然轻量：特征融合只增加线性拼接的开销，而自适应融合仅需一个小型元分类器。虽然我们的准确率在 80% 左右波动，但这与此前在含噪社交媒体文本上开展的最先进水平情感研究结果一致。推文和 Reddit帖子常包含讽刺、俚语和混合情绪，而标注者一致性本身也很少超过 85–90%。我们预计，如果将 LLaMA 或 Gemma 等更强大的模型纳入模型池，便能进一步提升各个基线模型的表现。不过，我们的框架与模型无关：即使大型模型也会误判含糊或带有讽刺意味的输入，融合仍能在这些情况下发挥作用。因此，SentiFuse 能够与大语言模型互补，而非与之竞争。未来的工作将探索更多上下文感知自适应机制，并将该框架扩展到多语言和领域特定场景。

资助

本研究获得美国国家科学基金会（美国）CNS-2219614资助。

参考文献

[1] Baccianella, S., Esuli, A., Sebastiani, F：SentiWordNet 3.0：一种增强型情感分析与观点挖掘词汇资源。载于：第七届国际语言资源与评估会议（LREC'10）论文集。第 2200–2204 页。欧洲语言资源协会（ELRA）（2010）[2] Behera, R.K., Jena, M., Rath, S.K., Misra, S：Co-LSTM：用于社交大数据情感分析的卷积 LSTM 模型。《信息处理与管理》**58(1)**，102435（2021）[3] Breiman, L：装袋预测器。《机器学习》**24(2)**，123–140（1996）[4] Chikersal, P., Poria, S., Cambria, E：Sentu：结合基于规则的分类器与监督学习进行推文情感分析。载于：Nakov, P., Zesch, T., Cer, D., Jurgens, D.（编）第九届语义评估国际研讨会（SemEval 2015）论文集。第 647–651 页（2015）[5] Chiny, M., Chihab, M., Bencharef, O., Chihab, Y：基于 LSTM、VADER 和词频-逆文档频率的混合情感分析模型。《国际高级计算机科学与应用期刊》**12(7)**（2021）[6] Das, B., Chakraborty, S：利用词频-逆文档频率和后续词否定改进文本情感分类模型（2018）[7] Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., Ravi, S：GoEmotions：一个细粒度情绪数据集。载于：Jurafsky, D., Chai, J., Schluter, N., Tetreault, J.（编）计算语言学协会第 58 届年会论文集。第 4040–4054 页。计算语言学协会，线上（2020）[8] Devlin, J., Chang, M.W., Lee, K., Toutanova, K：BERT：用于语言理解的深度双向 Transformer 预训练。载于：2019 年北美计算语言学协会分会：人类语言技术会议论文集，第 1 卷。第 4171–4186 页（2019）[9] Dey, R., Das, A：基于改进词频-逆文档频率的情感分析深度混合框架。《多媒体工具与应用》**82(21)**，32967–32990（2023）[10] Dietterich, T.G：机器学习中的集成方法。载于：多分类器系统国际研讨会。第 1–15 页。施普林格（2000）[11] Elbagir, S., Yang, J：使用自然语言工具包和 VADER 情感分析器进行推特情感分析。载于：国际工程师与计算机科学家多学科会议论文集。第 122 卷（2019）[12] Feldman, R：情感分析的技术与应用。《美国计算机协会通讯》**56(4)**，82–89（2013）。https://doi.org/10.1145/2436256.2436274[13] Gan, C., Fu, X., Feng, Q., Zhu, Q., Cao, Y., Zhu, Y：采用注意力机制的视觉-文本情感分析多模态融合网络。《应用专家系统》**242**，122731（2024）[14] Go, A., Bhayani, R., Huang, L：利用远程监督进行推特情感分类。技术报告，斯坦福大学（2009）

[15] Hutto, C., Gilbert, E：VADER：一种简洁的基于规则的社交媒体文本情感分析模型。载于：国际 AAAI 网络与社交媒体会议论文集。第 8 卷，第 216–225 页（2014）[16] Kiritchenko, S., Zhu, X., Mohammad, S.M：简短非正式文本的情感分析。《人工智能研究杂志》**50**，723–762（2014）[17] Kora, R., Mohammed, A：基于元集成深度学习的增强型情感分析方法。《社交网络分析与挖掘》**13**，38（2023）[18] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V：RoBERTa：一种稳健优化的 BERT 预训练方法（2019）[19] Lossio-Ventura, J.A., Weger, R., Lee, A.Y. 等：ChatGPT 与微调的开放式预训练变换器（OPT）和广泛使用的情感分析工具的比较：新冠肺炎调查数据的情感分析。《JMIR Mental Health》**11**，e50150（2024）[20] Miah, M. 等：结合 Transformer 模型与大语言模型的多语言情感分析集成模型。载于：国际自然语言处理会议（2024）[21] Mohammad, S.M., Kiritchenko, S., Zhu, X：NRC-Canada：打造推文情感分析的最先进水平。载于：第七届语义评估研讨会（SemEval-2013）论文集。第 321–327 页（2013）[22] Ravi, K., Ravi, V：观点挖掘与情感分析综述：任务、方法与应用。《基于知识的系统》**89**，14–46（2015）[23] Ray, P., Chakrabarti, A：结合深度学习方法与基于规则的方法，以改进方面级情感分析。《应用计算与信息学》**18(1/2)**，163–178（2022）[24] Reusens, M., Stevens, A., Tonglet, J., De Smedt, J., Verbeke, W., vanden Broucke, S., Baesens, B：文本分类评估：一项基准研究。《应用专家系统》**254**，124302（2024）[25] Rodríguez-Ibánez, M., Casánez-Ventura, A., Castejón-Mateos, F., Cuenca-Jiménez, P.M：社交媒体平台情感分析综述。《应用专家系统》**223**，119862（2023）[26] Salton, G., Buckley, C：自动文本检索中的词项加权方法。《信息处理与管理》**24(5)**，513–523（1988）[27] Singh, C., Imam, T., Wibowo, S., Grandhi, S：一种用于新冠肺炎评论情感分析的深度学习方法。《应用科学》**12(8)**，3709（2022）[28] Singh, U., Abhishek, K., Azad, H：多模态情感分析前沿综述。《ACM 计算调查》**56(9)**，1–38（2024）[29] Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M：基于词典的情感分析方法。《计算语言学》**37(2)**，267–307（2011）[30] Wang, L. 等：ChatGPT 是优秀的情感分析器吗？一项初步研究（2024）[31] Wang, X., Gan, Y：针对股票情绪预测模型的多级对抗训练。《社交网络分析与挖掘》（2023）[32] Xu, G., Meng, Y., Qiu, X., Yu, Z., Wu, X：基于双向长短期记忆网络的评论文本情感分析。IEEE Access**7**，51522–51532（2019）

[33] Yadav, A., Vishwakarma, D.K.：使用深度学习架构进行情感分析：综述。《人工智能评论》**53**(6), 4335–4385 (2020). https://doi.org/10.1007/s10462-019-09794- 5[34] Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H., Xu, B.：用于关系分类的基于注意力的双向长短期记忆网络。载于：第54届计算语言学协会年会论文集（第2卷：短篇论文）。第207–212页（2016）