认知与情感的桥梁：同理心驱动的多模态错误信息检测

Zihan Wang, Lu Yuan, Zhengxuan Zhang, Qing Zhao
Communication University of China
Beijing, China

Abstract

在数字时代，社交媒体已成为信息传播的主要渠道，但它也加速了错误信息的传播。传统的错误信息检测方法主要关注于表面特征，忽视了人类同理心在传播过程中的关键作用。为了解决这一问题，我们提出了双重视角同理心框架（DAE），该框架整合了认知和情感同理心，从创作者和读者两个角度分析错误信息。通过考察创作者的认知策略和情感诉求，以及利用大语言模型（LLMs）模拟读者的认知判断和情感反应，DAE 提供了一种更全面和以人为中心的错误信息检测方法。此外，我们进一步引入了一种考虑同理心的过滤机制，以增强反应的真实性和多样性。在基准数据集上的实验结果表明，DAE 优于现有的方法，为多模态错误信息检测提供了一种新的范式。

1 介绍

在数字时代的浪潮中，社交媒体平台已经成为信息交流的主要场所 [9, 23]，但它们也为错误信息的大规模传播提供了肥沃的土壤 [1, 34, 52, 57]。错误信息，如同数字瘟疫，以前所未有的速度和规模在网络空间传播，不仅扰乱了公众对真相的认知，而且深刻影响了社会共识的形成和公共政策的发展 [12, 42]。面对这一复杂而紧迫的挑战，建立高效且准确的错误信息识别机制已成为维护信息生态系统健康和社会稳定的关键任务 [25]。

近年来，虚假信息检测技术取得了显著进展，研究人员结合了文本分析 [7, 32]、用户行为建模 [33, 35]、图像识别 [41] 及其他技术 [17, 26]。许多研究也探讨了情感在虚假信息检测中的作用，利用情感信号提高分类准确性 [5, 20]。虚假信息的传播不仅是因为情感，也由于更深层次的心理和认知偏见。它不仅取决于内容本身，还取决于人们如何解读内容——这受创作者意图和读者偏见的影响 [6, 24, 59]。如图 1 所示，这种引导性虚假信息在日常生活中社交媒体上尤其常见。创作者的目的是影响读者的情感，引导他们点击和传播内容。因此，仅依赖于基于情感的内容分析可能会陷入创作者的圈套并相信“这是真的”。然而，如果同时从认知和情感的角度考虑，去理解内容和创作者，就更容易识别出这是假信息。

这一现象启发我们借鉴心理学中的共情理论来重建错误信息检测的理论框架。共情作为人类理解他人心理状态的核心能力，包含两个互补的维度：认知共情和情感共情 [31, 39]。认知共情使我们能够采纳他人的观点，理解他们的思维过程、信仰系统和行为意图，这对于分析错误信息制造者的策略（侧重于“是什么”）至关重要；而情感共情使我们能够感知和体验他人的情感状态，这在模拟错误信息对受众的情感影响方面具有独特价值（侧重于“如何”）。在错误信息检测中整合这两种共情能力不仅揭示了创造者更深层次的动机和策略，还模拟了不同接收者的情感和认知反应模式，从而构建一个更全面和人性化的检测框架。

基于这一理论，我们提出了一个双重视角共情框架（DAE），从创作者和读者的双重角度有机地结合认知和情感共情，为多模态错误信息检测提供了一个更深层次的分析路径。在创作者维度中，我们分析文本和视觉内容的特征以获得创作者的共情向量；在读者维度中，我们使用 LLMs 来模拟不同读者群体的情感反应和认知判断过程，评估信息的感知可信度，并

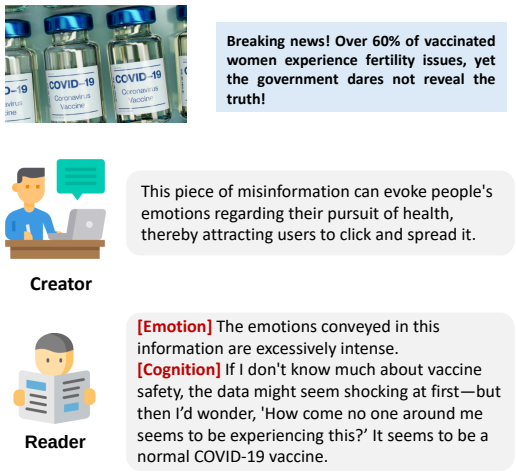


Figure 1: 一个错误信息的例子。

精心设计一个共情感知的过滤机制，以确保模拟反应的真实性和多样性。然后，我们在分析中将二者之间的差距作为一个特征纳入，以衡量其情感偏离。这个双重维度的框架可以捕捉错误信息传播的深层心理动力，突破传统方法的局限，实现更精确和更具解释力的检测效果。

我们的工作的主要贡献是：

- 我们提出了一种双重视角共情框架，这是一个基于创作者和读者视角的多层次分析，结合了认知和情感共情特征。
- 为了获得不同群体读者的共情特征，我们设计了一种创新的情感共情数据生成和过滤机制。
- 在两个基准数据集上进行了实验，结果表明我们的模型是检测多模态错误信息的有竞争力的方法。

2 相关工作

2.1 虚假信息检测

虚假信息是指传播的错误或误导性的信息，无论其是否有意欺骗 [1]。它对公众舆论、社会稳定和信息安全构成重大威胁 [56]。近年来，检测方法不断从传统的机器学习方法演变为深度学习 [3, 8, 57]。早期的虚假信息检测主要依赖于手工特征工程，通过提取文本语言特征和传播特征结合浅层分类模型进行预测 [4]。然而，这些方法在语义理解和上下文分析方面存在局限性，难以捕捉微妙的表达和复杂的上下文 [44]。

随着 LLM 的兴起，模型在捕捉远距离语义关联和逻辑推理方面取得了进步 [13]。最近的工作如 Dm-inter [46] 和 JSDRV [50] 通过推理特征和数据增强提升了检测能力。其他研究则利用用户评论作为辅助信息；例如，模拟评论 [26] 和语义一致性探索 [48]。

多模态学习进一步促进了文本-图像融合。像 att-RNN [15] 和 CAFE [10] 这样的模型推动了特征融合和模态一致性。在可

解释性方面，SNIFFER [28] 和 DELL [45] 改进了透明性和可靠性。

然而，上述研究忽略了虚假信息传播背后的心理因素，特别是创作者的情感操控和读者的情感-认知反应。为了解决这一空白，我们的方法独特地结合了来自创作者和读者角度的同理心理论与 LLM，整合认知和情感同理心分析，以更好地理解虚假信息的心理动态。

2.2 社会计算中的同理心

同理心，即理解和体验他人情感和认知状态的能力，在人工智能研究中正发挥着越来越重要的作用 [27, 40] 。研究人员正尝试将同理心功能纳入自然语言处理任务中，以促进在人机对话 [2] 以及心理健康 [18] 等应用领域的发展。

在虚假信息检测中，情感和共情视角提供了一种新的分析维度。Ma et al. [22] 通过易感于虚假信息测试模拟用户反应，而 [60] 则通过挖掘发布者情感与社交情感之间的关系来提高检测性能。这些研究表明，引入情感和共情视角有助于揭示虚假信息传播的深层心理机制，为传统的技术特征分析提供了有力的补充。

在评估和提升大型语言模型(LLMs)同理能力的过程中，Qian et al. [29] 和 Lee et al. [19] 的比较研究表明，LLMs 在同理对话生成方面已经展示出接近甚至超越人类的能力。此外，Zhu et al. [58] 和 Wang et al. [47] 的研究在算法和架构层面增强了 LLMs 的同理表达能力。然而，目前大多数研究往往只关注同理能力的单一维度，缺乏对同理心多维性质的系统考虑。

3 方法

在这一部分，我们首先介绍任务定义，随后是 DAE 框架概述。然后将详细描述每个组件，如图 2 所示。

3.1 任务定义

在这项工作中，我们特别关注基于新闻的错误信息。多模态错误信息检测任务旨在基于新闻的多模态输入信息确定给定新闻的真实性。形式上，每条新闻表示为：

$$D_i = \{T_i, I_i, C_i\}$$

其中 D_i 表示第 i 条新闻的多模态数据，T_i 代表文本内容，I_i 对应相关图片，C_i 是用户评论的集合。该任务的目标是将每条推文分类为真或假，使其成为一个二元分类问题：

$$f(D_i) \rightarrow \{\text{TRUE}, \text{FALSE}\}$$

其中，$f(\cdot)$ 代表分类模型，它将多模态数据 D_i 作为输入，预测推文是否包含虚假信息。

3.2 评论生成和过滤

为了捕捉读者对文字和视觉信息的“移情”反应，我们利用大型语言模型（LLMs）来模拟读者并自动生成评论。参考 Nan et al. [26] 的工作，不同的人口群体（如年龄、性别和教育背景）以不同的理解水平和视角感知内容。这些变化为我们关于共情的研究提供了有价值的见解，包括情感和认知的共情。因此，我们设计了一个基于 LLM 的评论生成和过滤机制，以全面捕捉读者的移情反应。

3.2.1 **基于人口统计的评论生成**. 增强过程始于使用 LLM 生成评论，以模拟多样的社会视角。通过指导 LLM 表现不同的人口背景，我们创建了一个更全面的真实世界公众话语的代表。我们定义了一组用户配置文件：

$$\text{Profiles} = \{P_1, P_2, P_3, \ldots, P_n\} \tag{1}$$

其中每个剖面 P_i 的特征是：
- 性别：男/女。
- 年龄组：青年 (18-35)，中年 (36-65)，老年 (65+)。
- 教育程度：学士以下、学士、研究生。

给定一个新闻文章 N 及其文本内容、相关图片 I 和任何现有评论集 $C = \{c_1, c_2, \ldots, c_m\}$ ，我们提示 GPT-4o 模拟来自不同背景人群的个人如何进行回应：

$$G = \text{LLM}(N, C, \text{Profiles}, I) \tag{2}$$

这种基于 LLM 的模拟方法使我们能够捕捉到不同个体可能如何处理、解释和响应潜在误导性多模态内容的细微方式。这种模拟由两个基本的移情维度引导：

认知同理心：我们指导大型语言模型去模拟来自不同背景的个人如何认知加工信息，包括他们识别不一致的能力、基于先前知识评估可信度，以及形成反映其教育和经验背景的观点的能力。如图 1 所示，具有医学或学术背景的人可能会立即质疑“60 %”的有效性，寻求证据并怀疑虚假信息。而其他人，尤其是科学素养较低的人，可能会先表现出惊慌，但在发现该声明与他们的日常经验相矛盾后，随即表达怀疑。

情感共鸣：LLM 模拟可能从不同人口群体中产生的情感反应，捕捉与年龄、性别和教育差异相关的情感强度、情感触发因素和情感表达的变化。如图 1 中的例子所示，年轻女性可能感到强烈的焦虑，尤其是关于未来家庭规划的问题。家长可能会感到愤怒或背叛，然后才会转向验证。

通过引导 LLM 体现这些移情维度，同时考虑到人口因素，我们获得了丰富的模拟反应谱，这反映了人类对多模态信息反应的复杂性和多样性。

3.2.2 **基于同理心的评论过滤**. 在模拟阶段之后，我们实施了一种稳健的过滤机制，以确保仅保留高质量、具有同理心的评论。我们获得了精炼的评论集 C' ，具体如下：

$$C' = \text{Filter}(C \cup G) \tag{3}$$

过滤标准优先考虑展示认知或情感共情的评论，同时排除与新闻主题无关的、缺乏共情信号、包含冒犯或不当内容、包括社交媒体元素（例如 @ 提及、标签）或使用非英语撰写的评论（这些评论已被翻译）。

为了保持情境的丰富性，我们对每篇文章强制执行至少五条高质量评论的最低门槛。对于评论数量过多的文章（$|C| > 50$ ），我们应用基于长度的优先排序，保留通常提供更多实质性见解的较长评论。

3.3 特征提取与编码

对于原始新闻的文本 T_i 和图像 I_i ，我们通过专用编码器提取特征：

$$T_i = \text{Encoder}_{\text{text}}(T_i), \tag{4}$$

$$I_i = \text{Encoder}_{\text{visual}}(I_i), \tag{5}$$

，其中 $T_i \in \mathbb{R}^{L_t \times d_t}$ 和 $I_i \in \mathbb{R}^{L_i \times d_i}$ 。为了确保特征空间的一致性，我们通过线性映射将文本和视觉特征映射到统一的维度

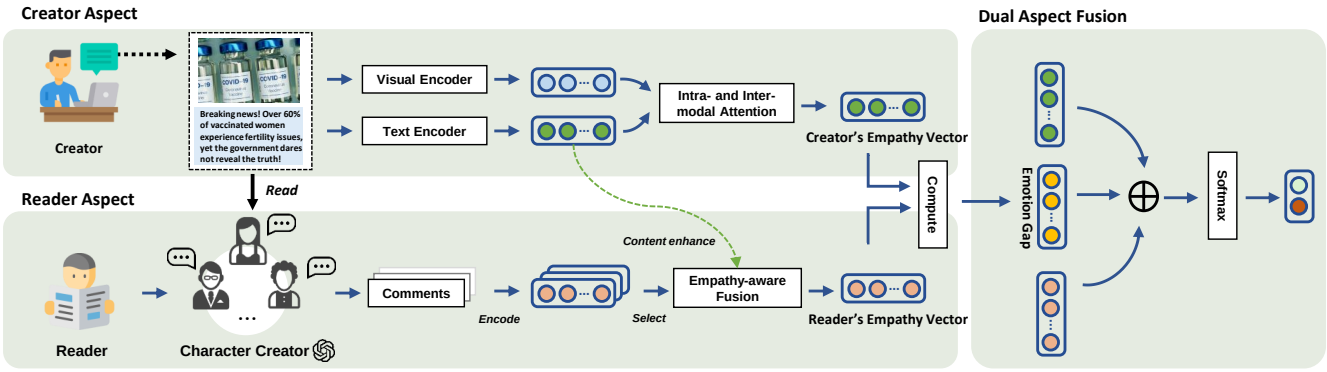


Figure 2: 我们提出的双重视角同理心框架 (DAE) 概述

空间：

$$T_i' = W_t T_i, \tag{6}$$

$$I_i' = W_i I_i, \tag{7}$$

，其中 $T_i' \in \mathbb{R}^{L_t \times d}$ 和 $I_i' \in \mathbb{R}^{L_i \times d}$ 。与仅关注内容本身的传统方法不同，我们特别引入用户评论作为读者观点的重要补充。对于相关的评论集合 C_i ，我们为每条评论提取语义表示：

$$C_{ij} = \text{Encoder}_{\text{text}}(C_{ij})_{[\text{CLS}]}, \tag{8}$$

，其中 [CLS] 表示 RoBERTa 输出的分类标记向量，并且 $C_{ij} \in \mathbb{R}^{d_t}$ 。所有评论形成集体表示：

$$C_i = [C_{i1}; C_{i2}; \ldots; C_{iN}], \tag{9}$$

，其中 $C_i \in \mathbb{R}^{N \times d_t}$ 。类似地，我们对评论特征应用线性降维：

$$C_i' = W_c C_i, \tag{10}$$

，使用 $C_i' \in \mathbb{R}^{N \times d}$ 。

3.4 特征增强和交互

为了捕捉序列内部依赖关系，我们将多头自注意力机制（MHSA）分别应用于文本、图像和评论特征：

$$T_i'' = \text{MHSA}(T_i'), I_i'' = \text{MHSA}(I_i'), C_i'' = \text{MHSA}(C_i') \tag{11}$$

通过跨模态注意力机制，我们融合了文本和视觉特征：

$$TI_i = \text{CrossAttn}(T_i'', I_i'') \tag{12}$$

对于评论收集，我们采用一种基于语义重要性的自适应过滤机制：

$$C_i^{\text{k}} = \text{top-k}(C_i'', k) \tag{13}$$

这确保了被选中的评论最大限度地反映公众对信息的真实反应，提供更丰富的读者视角证据。

3.5 双视角同理特征构建

3.5.1 **基本特征提取** . 我们首先使用池化操作提取两种基本类型的特征。具体来说，我们应用均值池化获得创作者特征，反映信息生成者的整体意图和认知模式。相反，最大池化用于获取读者特征，捕捉观众最显著的反应：

$$e_{\text{creator}} = \text{MeanPooling}(T_i''), e_{\text{reader}} = \text{MaxPooling}(C_i^{\text{k}}) \tag{14}$$

3.5.2 **同理心感知融合** . 为建模同理心的认知方面，我们将提取的创造者和读者特征与从文本和图像中获得的多模态表示相结合。这导致一种认知上知情的融合表示：

$$h_c = \text{Concat}(e_{\text{creator}}, e_{\text{reader}}, TI_i) \tag{15}$$

这种融合使模型能够共同捕捉创作者的沟通意图、读者的解释立场以及文本和视觉模式之间的语义对齐（或不对齐）。它反映了对信息如何构建和被感知的整体理解。

3.5.3 **情感差距计算** . 超越事实分析后，我们引入情感同理心维度，以更好地理解常见于虚假信息中的情感操控机制。第一步是计算创作者和读者之间的情感差异：

$$e_{\text{gap}} = |e_{\text{creator}} - e_{\text{reader}}| \tag{16}$$

这种情感差距捕捉了情感煽动的关键信号，其中创作者可能会故意夸大或扭曲情感线索以影响读者的感知。然后我们通过融合双方的情感信号及其差异构建一个综合的情感共鸣表示：

$$h_e = \text{Concat}(e_{\text{creator}}, e_{\text{reader}}, e_{\text{gap}}) \tag{17}$$

3.6 融合与分类

为了形成一个统一的表示，我们合并认知和情感共情特征：

$$h_f = \text{Concat}(h_c, h_e) \tag{18}$$

最终的表示通过多层感知机 (MLP)，然后通过 Softmax 激活来预测内容的真实性：

$$\hat{y}_i = \text{Softmax}(\text{MLP}(h_f)) \tag{19}$$

对于训练，我们使用带有标签平滑的交叉熵损失函数，结合 L_2 正则化来防止过拟合：

$$L(\theta) = -\sum_i \sum_c y_{ic} \log \hat{y}_{ic} + \lambda \|\theta\|^2 \tag{20}$$

其中 y_{ic} 表示真实的标签分布，λ 是正则系数。

4 实验

4.1 实验设置

数据集。我们在两个广泛使用的数据集上评估模型性能：PHEME [54] 和 PolitiFact [14] 。PHEME 数据集是一个多模态谣传集合，包括五个突发事件：查理周刊事件、弗格森事件、德国之翼航空失事、渥太华枪击案和悉尼人质事件。我们对与每个新闻相关的推文、图像和评论进行分类。PolitiFact 也是一个

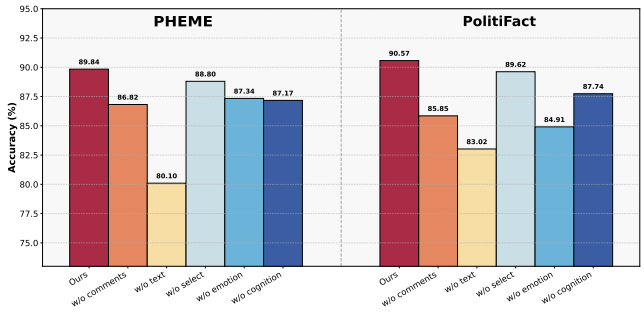


Figure 3: 消融研究结果

多模态假新闻检测数据集。两个数据集中真实新闻和假新闻的分布如表所示，我们采用 8 : 2 比例进行训练集和测试集的拆分。

Table 1: 划分两个数据集的数据集

Partition	Datasets	
	PHEME	PolitiFact
# of Fake News	1972	432
# of True News	3830	624
# of Images	3670	783

评价指标。我们使用准确率、精确率、召回率和 F1-分数作为评价指标。

实验环境。在我们的实验中，最大文本序列长度设置为 512 ，每个评论序列的最大长度为 128 。由于与新闻条目相关的评论数量有所不同，我们将每个新闻条目限制为最多 15 条真实评论，并引入 GPT 生成的补充评论。对于文本编码，我们使用预训练的 RoBERTa-base 模型，而视觉编码采用预训练的 Swin-base 模型，冻结 Swin Transformer 的前 $n-1$ 层，并且仅训练最后一层。多头自注意力网络使用 8 个注意力头。图像特征通过 Swin Transformer 提取，输出序列长度由 Swin 模型的设置决定。在评论选择机制中，Pointer Network 选择排名前 k 的评论，其中 $k = 5$ 。此外，初始学习率为 0.001 ，优化器是 Adam，损失函数是交叉熵损失。

4.2 基准线

为了验证我们模型的性能，我们选择了几种具有竞争力的错误信息检测方法作为基线。对于 PHEME 数据集，基线方法是 MVAE [16] 、SAFE [55] 、SpotFake [38] 、CAFE [10] 、MCAN [49] 、KDIN [43] 、LIIMR [37] 、BMR [53] ，而对于 PolitiFact，其基线包括 XLNet [51] 、GCAN [21] 、SpotFake [38] 、KAN [11] 、MCAN [49] 、MCAN-A [49] 、SpotFake+ [36] 、QMFND [30] 。

4.3 主要实验

为了验证我们方法的有效性，我们在两个数据集上进行了实验。主要的实验结果如表 2 所示。总体而言，我们的方法在两个数据集上都优于基线。

具体来说，在 PHEME 数据集上，我们提出的 DAE 模型取得了 89.8 % 的准确率，显著超过了最新的 BMR（88.4 %）和 LIIMR（87.0 %）。详细而言，BMR 利用一种双向多模态推理机制，能够深入捕捉文本和视觉信息之间的语义交互，而 LIIMR 则通过隐式交互推断探索多模态特征之间的潜在关系。然而，这两种方法主要关注内容本身，并未明确考虑用户在情感和认知维度上的反馈。相比之下，DAE 不仅整合了多模态信息，包括

Table 2: 主要实验结果

Dataset	Method	Acc(%)	P.(%)	R.(%)	F1(%)
PHEME	MVAE	77.6	73.5	72.3	72.8
	SAFE	80.7	78.7	78.9	79.1
	SpotFake	84.5	80.9	83.6	82.2
	CAFE	83.2	79.6	79.4	79.5
	MCAN	86.1	83.0	84.0	83.5
	KDIN	84.6	81.5	80.4	80.9
	LIIMR	87.0	84.8	83.1	83.9
	BMR	88.4	87.2	84.0	85.5
	DAE (Ours)	89.8	89.9	89.8	89.8
PolitiFact	XLNet	84.7	90.5	70.4	79.2
	GCAN	80.8	79.5	84.2	83.5
	SpotFake	77.9	81.5	69.3	74.9
	KAN	85.9	86.9	85.0	85.4
	MCAN	84.6	85.1	82.9	84.0
	MCAN-A	80.9	82.7	75.8	79.2
	SpotFake+	78.9	80.3	75.3	77.7
	QMFND	84.6	92.7	85.3	88.8
	DAE (Ours)	90.6	90.8	90.6	90.4

文本、图片和评论，还通过双维度共鸣机制明确建模用户的情感共鸣和认知怀疑。这种综合建模更好地反映了用户对新闻的真实反馈，使模型能够更准确地识别新闻的真实性。

在 PolitiFact 数据集上，DAE 也实现了显著的提升，准确率达到 90.6%，明显优于采用基于核的注意力机制的 KAN（85.9%）和通过量子卷积神经网络整合图像和文本特征的 QMFND（84.6%）。KAN 主要通过基于核的注意力来测量文本语义相似性，能够捕捉丰富的语义信息，但忽略了用户评论与新闻内容之间的互动关系。尽管 QMFND 融合了文本和视觉特征，但其采用静态特征提取的方法，未能充分利用用户反馈中蕴含的情感和认知线索。相比之下，DAE 包含了评论选择机制，能够明确捕捉用户对新闻内容的认知一致性和情感分歧。这使得模型能够深入探索用户对新闻的共鸣或质疑。

4.4 消融研究

我们通过五种消融设置验证了双维度共情框架中每个组件的有效性，包括：

- w/o 评论：去除读者评论输入的 DAE。
- w/o 文本：移除推文文本输入的自编码器（DAE）。
- w/o 选择：去除注释选择机制的 DAE。
- w/o 情感：去除情感共鸣的 DAE。
- w/o 认知：移除认知共情的 DAE。

实验结果表明，推文文本是基础，因为移除它会导致最大的准确性下降（PHEME: 从 89.84 % 降至 80.10 % ； PolitiFact: 从 90.57 % 降至 83.02 % ）。移除评论也会降低性能（PHEME: -3.02 % ， PolitiFact: -4.72 % ），这确认了评论的互补作用。

评论选择机制显著提高了准确性，因为其移除会降低性能（PHEME： -1.04 % ， PolitiFact： -0.95 % ），验证了 top-k 过滤的有效性。

情感移情和认知移情对数据集的影响不同。在 PolitiFact 数据集上，去除情感移情导致的下降更大（从 90.57 % 下降到 84.91 %），而去除认知移情的影响较小（从 90.57 % 下降到 87.74 %）。相反，在 PHEME 数据集上，认知移情的影响更大（从

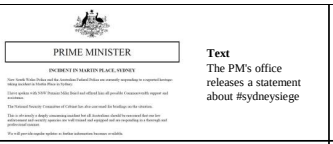
Samples	 **Text** #RIGHTNOW all bridges from #Ottawa #Gatineau are NOW CLOSED!	 **Text** Map shows industrial estate where 2 #CharlieHebdo suspects are holed up, surrounded by police	 PRIME MINISTER INCIDENT IN MARTIN PLACE, SYDNEY **Text** The PM's office releases a statement about #sydneysiege	 **Text** URGENT: Both #CharlieHebdo suspects killed
Real Comments	May The Lord grant the capture and punishment of these terrorists who cowardly attack our beloved neighbor, Canada!.....	To BBC Breaking : suggest you change 'martyr' themselves to 'kill themselves' in your TV news text - don't dignify them.....	The PM does not speak to nation - releases a press statement!..... but benefit of the doubt here: he probably has life-saving decisions to make??	Being kept alive would have been preferable. I do not trust this story so far, there are some big questions and failures.....
Generated Comments	It's important to maintain public safety, **but shutting down all bridges seems quite drastic**.....	While ....., the media's portrayal **often sensationalizes these events**. We should remain **calm** and trust law enforcement to handle the situation professionally without unnecessary dramatization.....	It's important for the PM to address the nation directly during such critical events. **A press statement feels impersonal and doesn't provide the reassurance people need at this time**.....	I have **serious reservations about the narrative being presented**. We've seen similar stories before, and they often raise more questions than answers.....
Ground Truth	FAKE	FAKE	FAKE	FAKE
w/o text	√	×	×	×
w/o select	×	×	×	×
w/o comments	×	×	√	√
w/o cognition	√	×	×	√
w/o emotion	×	×	×	√
Ours	√	√	√	√

Figure 4: 数据集测试集中的几种情况

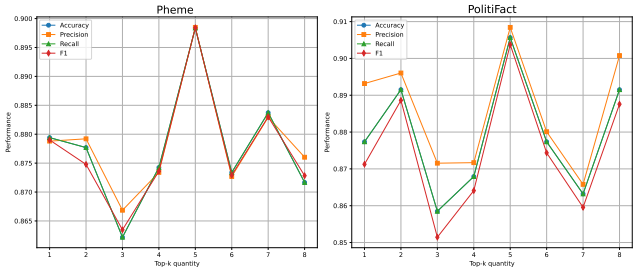


Figure 5: 参数分析结果

89.84 % 下降到 87.17 %），相比之下，去除情感移情的影响较小（从 89.84 % 下降到 87.34 %）。

这些结果突出了推文文本、读者评论、评论选择以及双维度同理心机制在模型性能中的必要性。

4.5 参数分析

在参数实验中，我们重点考察了 top-k 评论选择机制中的选择数量。我们在 PHEME 和 PolitiFact 数据集上测试了从 1 到 8 的不同评论选择数量，结果如图所示。实验发现，当选择的评论数量过少（如 1 或 2）时，模型无法获得足够的信息来支持决策；而当选择的评论数量过多（如 7 或 8）时，模型容易受到不相关或冗余评论的干扰，性能也会下降。

在这两个数据集中，当评论选择数量设置为 5 时，模型实现了最佳性能（在 PHEME 数据集上的准确率为 89.8 %，而在 PolitiFact 数据集上为 90.6 %）。这表明，适度数量的精选评论可以更有效地帮助模型集中于真正有价值的评论信息，从而提高虚假信息检测的准确性。

4.6 案例研究

如图 4 所示，我们进一步验证了双维度同理心框架的有效性以及每个模块的重要性。我们的研究发现，情感和认知同理心在评估新闻真实性中起着至关重要的作用。

在第一种情况下，真实评论传达了愤怒，而 AI 生成的评论则保持理性但具有批判性。移除情感共鸣显著降低了模型的准确性，凸显了其重要性。在第二种情况下，真实评论更具攻击性，而 AI 生成的评论通过批评媒体夸大其词来促进理性。认知共鸣的移除显著影响了模型的判断，强调了其重要性。在第三种情况下，真实评论批评政府的响应，而 AI 生成的评论则强调直接沟通。缺少用户评论导致检测准确性急剧下降，证明了其关键作用。在最后一种情况下，真实评论表达了不信任，而 AI 生成的评论则质疑官方的说法。移除认知和情感共鸣都导致了显著的误判，重申了其综合作用。

总体而言，我们的研究为情感和认知同理心在识别错误信息中的不可或缺作用提供了有力的实证证据，展示了它们在理论上的相互依赖性。

5 结论

总之，我们引入了一个用于多模态虚假信息检测的双重视角同理心框架，该框架整合了认知和情感同理心，以提供更深入、更以人为本的方法。通过分析创作者的心理策略并模拟读者的情感和认知反应，该框架相比传统方法实现了更为精确和具解释性的检测。

未来方向可以集中于实时同理心建模，以适应不断发展的错误信息策略，并将框架扩展到跨文化背景，以提高鲁棒性。

References

[1] Sara Abdali, Sina Shaham, and Bhaskar Krishnamachari. 2024. Multi-modal misinformation detection: Approaches, challenges and opportunities. *Comput. Surveys* 57, 3 (2024), 1–29.

[2] Firoj Alam, Morena Danieli, and Giuseppe Riccardi. 2018. Annotating and modeling empathy in spoken conversations. *Computer Speech & Language* 50 (2018), 40–61.

[3] Jawaher Alghamdi, Suhuai Luo, and Yuqing Lin. 2024. A comprehensive survey on machine learning approaches for fake news detection. *Multimedia Tools and Applications* 83, 17 (2024), 51009–51067.

[4] Mohammad Q Alnabhan and Paula Branco. 2024. Fake news detection using deep learning: A systematic literature review. *IEEE Access* (2024).

[5] Miguel A Alonso, David Vilares, Carlos Gómez-Rodríguez, and Jesús Vilares. 2021. Sentiment analysis for fake news detection. *Electronics* 10, 11 (2021), 1348.

[6] Vian Bakir and Andrew McStay. 2018. Fake news and the economy of emotions: Problems, causes, solutions. *Digital journalism* 6, 2 (2018), 154–175.

[7] Waqas Haider Bangyal, Rukhma Qasim, Najeeb Ur Rehman, Zeeshan Ahmad, Hafsa Dar, Laiqa Rukhsar, Zahra Aman, and Jamil Ahmad. 2021. Detection of Fake News Text Classification on COVID-19 Using Deep Learning Approaches. *Computational and mathematical methods in medicine* 2021, 1 (2021), 5514220.

[8] Elena Broda and Jesper Strömbäck. 2024. Misinformation, disinformation, and fake news: lessons from an interdisciplinary, systematic literature review. *Annals*

of the International Communication Association 48, 2 (2024), 139–166.

[9] Sijing Chen, Lu Xiao, and Akit Kumar. 2023. Spread of misinformation on social media: What contributes to it and how to combat it. *Computers in Human Behavior* 141 (2023), 107643.

[10] Yixuan Chen, Dongsheng Li, Peng Zhang, Jie Sui, Qin Lv, Lu Tun, and Li Shang. 2022. Cross-modal ambiguity learning for multimodal fake news detection. In *Proceedings of the ACM web conference 2022*. 2897–2905.

[11] Yaqian Dun, Kefei Tu, Chen Chen, Chunyan Hou, and Xiaojie Yuan. 2021. Kan: Knowledge-aware attention network for fake news detection. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 81–89.

[12] Bin Guo, Yasan Ding, Lina Yao, Yunji Liang, and Zhiwen Yu. 2019. The future of misinformation detection: new perspectives and trends. *arXiv preprint arXiv:1909.03654* (2019).

[13] Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. 2024. Bad actor, good advisor: Exploring the role of large language models in fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 22105–22113.

[14] Yiqiao Jin, Xiting Wang, Ruichao Yang, Yizhou Sun, Wei Wang, Hao Liao, and Xing Xie. 2022. Towards fine-grained reasoning for fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 5746–5754.

[15] Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. 2017. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM international conference on Multimedia*. 795–816.

[16] Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. Mvae: Multimodal variational autoencoder for fake news detection. In *The world wide web conference*. 2915–2921.

[17] Jooyeon Kim, Behzad Tabibian, Alice Oh, Bernhard Schölkopf, and Manuel Gomez-Rodriguez. 2018. Leveraging the crowd to detect and reduce the spread of fake news and misinformation. In *Proceedings of the eleventh ACM international conference on web search and data mining*. 324–332.

[18] Yoon Kyung Lee, Inju Lee, Minjung Shin, Seoyeon Bae, and Sowon Hahn. 2023. Chain of empathy: Enhancing empathetic response of large language models based on psychotherapy models. *arXiv preprint arXiv:2311.04915* (2023).

[19] Yoon Kyung Lee, Jina Suh, Hongli Zhan, Junyi Jessy Li, and Desmond C Ong. 2024. Large language models produce responses perceived to be empathic. *arXiv preprint arXiv:2403.18148* (2024).

[20] Zhiwei Liu, Tianlin Zhang, Kailai Yang, Paul Thompson, Zeping Yu, and Sophia Ananiadou. 2024. Emotion detection for misinformation: A review. *Information Fusion* 107 (2024), 102300.

[21] Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting rumors from microblogs with recurrent neural networks. (2016).

[22] Weicheng Ma, Chunyuan Deng, Aram Moossavi, Lili Wang, Soroush Vosoughi, and Diyi Yang. 2024. Simulated misinformation susceptibility (smists): Enhancing misinformation research with large language model simulations. In *Findings of the Association for Computational Linguistics ACL 2024*. 2774–2788.

[23] Weixing Mai, Zhengxuan Zhang, Kuntao Li, Yun Xue, and Fenghuan Li. 2023. Dynamic graph construction framework for multimodal named entity recognition in social media. *IEEE Transactions on Computational Social Systems* 11, 2 (2023), 2513–2522.

[24] Cameron Martel, Gordon Pennycook, and David G Rand. 2020. Reliance on emotion promotes belief in fake news. *Cognitive research: principles and implications* 5 (2020), 1–20.

[25] Muhammad Firoz Mridha, Ashfia Jannat Keya, Md Abdul Hamid, Muhammad Mostafa Monowar, and Md Saifur Rahman. 2021. A comprehensive review on fake news detection with deep learning. *IEEE access* 9 (2021), 156151–156170.

[26] Qiong Nan, Qiang Sheng, Juan Cao, Beizhe Hu, Danding Wang, and Jintao Li. 2024. Let silence speak: Enhancing fake news detection with generated comments from large language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 1732–1742.

[27] Anat Perry. 2023. AI will never convey the essence of human empathy. *Nature Human Behaviour* 7, 11 (2023), 1808–1809.

[28] Peng Qi, Zehong Yan, Wynne Hsu, and Mong Li Lee. 2024. Sniffer: Multimodal large language model for explainable out-of-context misinformation detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 13052–13062.

[29] Yushan Qian, Wei-Nan Zhang, and Ting Liu. 2023. Harnessing the power of large language models for empathetic response generation: Empirical investigations and improvements. *arXiv preprint arXiv:2310.05140* (2023).

[30] Zhiguo Qu, Yunyi Meng, Ghulam Muhammad, and Prayag Tiwari. 2024. QMFND: A quantum multimodal fusion-based fake news detection model for social media. *Information Fusion* 104 (2024), 102172.

[31] Katherine P Rankin, Joel H Kramer, and Bruce L Miller. 2005. Patterns of cognitive and emotional empathy in frontotemporal lobar degeneration. *Cognitive and Behavioral Neurology* 18, 1 (2005), 28–36.

[32] Harita Reddy, Namratha Raj, Manali Gala, and Annappa Basava. 2020. Text-mining-based fake news detection using ensemble methods. *International journal of automation and computing* 17, 2 (2020), 210–221.

[33] Natali Ruchansky, Sungyong Seo, and Yan Liu. 2017. Csi: A hybrid deep model for fake news detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 797–806.

[34] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter* 19, 1 (2017), 22–36.

[35] Kai Shu, Xinyi Zhou, Suhang Wang, Reza Zafarani, and Huan Liu. 2019. The role of user profiles for fake news detection. In *Proceedings of the 2019 IEEE/ACM international conference on advances in social networks analysis and mining*. 436–439.

[36] Shivangi Singhal, Anubha Kabra, Mohit Sharma, Rajiv Ratn Shah, Tanmoy Chakraborty, and Ponnurangam Kumaraguru. 2020. Spotfake+: A multimodal framework for fake news detection via transfer learning (student abstract). In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 13915–13916.

[37] Shivangi Singhal, Tanisha Pandey, Saksham Mrig, Rajiv Ratn Shah, and Ponnurangam Kumaraguru. 2022. Leveraging intra and inter modality relationship for multimodal fake news detection. In *Companion Proceedings of the Web Conference 2022*. 726–734.

[38] Shivangi Singhal, Rajiv Ratn Shah, Tanmoy Chakraborty, Ponnurangam Kumaraguru, and Shin'ichi Satoh. 2019. Spotfake: A multi-modal framework for fake news detection. In *2019 IEEE fifth international conference on multimedia big data (BigMM)*. IEEE, 39–47.

[39] Adam Smith. 2006. Cognitive empathy and emotional empathy in human behavior and evolution. *The Psychological Record* 56, 1 (2006), 3–21.

[40] Vera Sorin, Dana Brin, Yiftach Barash, Eli Konen, Alexander Charney, Girish Nadkarni, and Eyal Klang. 2024. Large Language Models and Empathy: Systematic Review. *Journal of Medical Internet Research* 26 (2024), e52597.

[41] Martin Steinebach, Karol Gotkowski, and Hujian Liu. 2019. Fake news detection by image montage recognition. In *Proceedings of the 14th international conference on availability, reliability and security*. 1–9.

[42] Qi Su, Mingyu Wan, Xiaoqian Liu, and Chu-Ren Huang. 2020. Motivations, methods and metrics of misinformation detection: an NLP perspective. *Natural Language Processing Research* 1, 1 (2020), 1–13.

[43] Mengzhu Sun, Xi Zhang, Jianqiang Ma, Sihong Xie, Yazheng Liu, and Philip S Yu. 2023. Inconsistent matters: A knowledge-guided dual-consistency network for multi-modal rumor detection. *IEEE Transactions on Knowledge and Data Engineering* 35, 12 (2023), 12736–12749.

[44] Michail Tsikerdekis and Sherali Zeadally. 2023. Misinformation Detection Using Deep Learning. *IT Professional* 25, 5 (2023), 57–63. doi:10.1109/MITP.2023.3314752

[45] Herun Wan, Shangbin Feng, Zhaoxuan Tan, Heng Wang, Yulia Tsvetkov, and Minnan Luo. 2024. Dell: Generating reactions and explanations for llm-based misinformation detection. *arXiv preprint arXiv:2402.10426* (2024).

[46] Bing Wang, Ximing Li, Changchun Li, Bo Fu, Songwen Pei, and Shengsheng Wang. 2024. Why Misinformation is Created? Detecting them by Integrating Intent Features. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 2304–2314.

[47] Lanrui Wang, Jiangnan Li, Zheng Lin, Fandong Meng, Chenxu Yang, Weiping Wang, and Jie Zhou. 2022. Empathetic dialogue generation via sensitive emotion recognition and sensible knowledge selection. *arXiv preprint arXiv:2210.11715* (2022).

[48] Lianwei Wu, Pusheng Liu, Yongqiang Zhao, Peng Wang, and Yangning Zhang. 2023. Human cognition-based consistency inference networks for multi-modal fake news detection. *IEEE Transactions on Knowledge and Data Engineering* 36, 1 (2023), 211–225.

[49] Yang Wu, Pengwei Zhan, Yunjian Zhang, Liming Wang, and Zhen Xu. 2021. Multimodal fusion with co-attention networks for fake news detection. In *Findings of the association for computational linguistics: ACL-IJCNLP 2021*. 2560–2569.

[50] Ruichao Yang, Wei Gao, Jing Ma, Hongzhan Lin, and Bo Wang. 2024. Reinforcement tuning for detecting stances and debunking rumors jointly with large language models. *arXiv preprint arXiv:2406.02143* (2024).

[51] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems* 32 (2019).

[52] WU Yin, Zhengxuan Zhang, WANG Fuling, Yuyu Luo, Hui Xiong, and Nan Tang. [n. d.]. Detecting Out-of-Context Misinformation via Multi-Agent and Multi-Grained Retrieval. ([n. d.]).

[53] Qichao Ying, Xiaoxiao Hu, Yangming Zhou, Zhenxing Qian, Dan Zeng, and Shiming Ge. 2023. Bootstrapping multi-view representations for fake news detection. In *Proceedings of the AAAI conference on Artificial Intelligence*, Vol. 37. 5384–5392.

[54] Qiang Zhang, Jiawei Liu, Fanrui Zhang, Jingyi Xie, and Zheng-Jun Zha. 2024. Natural language-centered inference network for multi-modal fake news detection. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*. 2542–2550.

[55] Xinyi Zhou, Jindi Wu, and Reza Zafarani. 2020. : Similarity-aware multi-modal fake news detection. In *Pacific-Asia Conference on knowledge discovery and data mining*. Springer, 354–367.

[56] Xinyi Zhou and Reza Zafarani. 2019. Fake news detection: An interdisciplinary research. In *Companion proceedings of the 2019 world wide web conference*. 1292–1292.

[57] Xinyi Zhou and Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)* 53, 5 (2020), 1–40.

[58] Jiahao Zhu, Zijian Jiang, Boyu Zhou, Jionglong Su, Jiaming Zhang, and Zhihao Li. 2024. Empathizing Before Generation: A Double-Layered Framework for Emotional Support LLM. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*. Springer, 490–503.

[59] Fabiana Zollo, Petra Kralj Novak, Michela Del Vicario, Alessandro Bessi, Igor Mozetič, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. 2015. Emotional dynamics in the age of misinformation. *PloS one* 10, 9 (2015), e0138740.

[60] Arkaitz Zubiaga, Maria Liakata, and Rob Procter. 2017. Exploiting context for rumour detection in social media. In *Social Informatics: 9th International Conference, SocInfo 2017, Oxford, UK, September 13-15, 2017, Proceedings, Part I 9*. Springer, 109–123.