# 基于证据生成和语义融合的虚假新闻检测

马 霄[1)] 黎欣雨[1)] 曾江峰[2)]

[1)](中南财经政法大学信息工程学院 武汉 430073)

[2)](华中师范大学信息管理学院 武汉 430079)

**摘 要** 虚假新闻检测作为自动化辅助手段,能够降低人工审核成本,实现内容审核流程的智能化升级,为构建高效可靠的信息真实性保障体系提供重要技术支撑。当前基于内容的虚假新闻检测方法主要聚焦于新闻文本内容的语义理解,缺乏对新闻真实性判定背后证据的深度挖掘,导致检测精度不足。尽管已有研究通过外部证据检索增强检测效果,但由此引入的噪声影响了检测结果的准确性和可靠性。近年来,大语言模型展现出卓越的语言理解能力,为证据发现提供了新的路径,能够有效避免外部证据检索带来的噪声问题。鉴于此,秉承大小语言模型协同的思想,本文提出了一种基于证据生成和语义融合的两阶段虚假新闻检测方法CollabDetection,探索如何利用大语言模型稳定生成高质量的潜在证据,并微调预训练小语言模型,对证据和新闻文本内容进行编码和特征融合,充分发挥大小语言模型各自的优势。在证据生成阶段,首先定义了涵盖信息准确性、情境关联性、信息来源可靠性等七个证据评估维度;其次,引入了样本级自适应上下文构造策略,动态选取示例并优化排序,为大语言模型生成有序且高质量的上下文语境;最后,模拟人类多专家决策过程,设计了基于生成对抗思想的多大语言模型协商框架,通过多轮对抗协商生成全面可靠的多维度证据,以有效降低由幻觉问题引入的噪声。在语义融合阶段,微调RoBERTa,提取新闻文本内容和多维度证据特征,通过多模态因子分解双线性池化技术实现特征间的充分融合,提升信息聚合效率。在Twitter15、Weibo16和Pheme9三个真实数据集上的大量实验证明,CollabDetection显著优于基于内容语义和特征融合的基线模型,Twitter15数据集上$F1$值提升了2.8%到46.6%,Weibo16数据集上提升了0.3%到34.2%,Pheme9数据集上提升了3.3%到59.3%。

**关键词** 虚假新闻检测;证据生成;语义融合;对抗协商;自适应上下文构造

**中图法分类号** TP18 **DOI号** 10.11897/SP.J.1016.2026.00365

# Fake News Detection Based on Evidence Generation and Semantic Fusion

MA Xiao[1)] LI Xin-Yu[1)] ZENG Jiang-Feng[2)]

[1)](School of Information Engineering, Zhongnan University of Economics and Law, Wuhan 430073)

[2)](School of Information Management, Central China Normal University, Wuhan 430079)

**Abstract** Fake news detection, as an automated auxiliary tool, can reduce the cost of manual review, enable intelligent upgrades to content moderation processes, and provide crucial technical support for building an efficient and reliable information authenticity assurance system. However, current content-based fake news detection methods mainly focus on semantic understanding of the news text, lacking in-depth exploration of the evidence behind the authenticity judgment, which leads to insufficient detection accuracy. Although some studies have improved detection performance through external evidence retrieval, the introduced noise has affected the accuracy and reliability of results. In recent years, large language models (LLMs) have demonstrated

收稿日期:2025-03-14;在线发布日期:2025-10-23。本课题得到国家自然科学基金(No. 62102159)、湖北省自然科学基金(No. 2024AFB957和No. 2023AFB1018)、教育部人文社会科学研究青年基金项目(21YJC870002)资助。**马 霄**,博士,副教授,中国计算机学会(CCF)会员,主要研究领域为推荐系统、虚假新闻检测、图表示学习、数据挖掘。E-mail:cindyma@zuel.edu.cn。**黎欣雨**,硕士研究生,主要研究领域为虚假新闻检测。**曾江峰**(通信作者),博士,副教授,中国计算机学会(CCF)会员,主要研究领域为虚假新闻检测、情感计算、推荐系统、人工智能。E-mail:jfzeng@ccnu.edu.cn。

exceptional language understanding capabilities, offering new avenues for evidence discovery and effectively mitigating the noise problems introduced by external retrieval. In this context, following the idea of collaboration between large and small language models, this paper proposes a two-stage fake news detection method based on evidence generation and semantic fusion, named CollabDetection. It explores how to leverage LLMs to stably generate high-quality potential evidence and fine-tune pre-trained small language models to encode and integrate the features of both the evidence and news content, thus fully utilizing the strengths of both types of models. In the evidence generation stage, the method first defines seven evidence evaluation dimensions, including information accuracy, contextual relevance, and source reliability. Then, a sample-level adaptive context construction strategy is introduced to dynamically select and optimally order examples, enabling LLMs to generate ordered and high-quality contextual environments. Finally, simulating the human multi-expert decision-making process, a generative adversarial negotiation framework involving multiple LLMs is designed to generate comprehensive and reliable multi-dimensional evidence through multi-round adversarial negotiation, effectively reducing noise caused by hallucinations. In the semantic fusion stage, RoBERTa is fine-tuned to extract features from the news content and the generated multi-dimensional evidence. These features are then fully fused using multimodal factorized bilinear pooling to improve the efficiency of information aggregation. Extensive experiments on three real-world datasets (i. e. , Twitter15, Weibo16 and Pheme9) demonstrate that CollabDetection significantly outperforms the baseline models based on content semantics and feature fusion. The $F1$ score improved by 2.8% to 46.6% on the Twitter15 dataset, by 0.3% to 34.2% on the Weibo16 dataset, and by 3.3% to 59.3% on the Pheme9 dataset.

**Keywords** fake news detection; evidence generation; semantic fusion; adversarial negotiation; adaptive in-context construction

## 1 引 言

互联网的深度普及加速了“信息时代”的到来，数字信息的传播形式和内容呈现多元化。但与此同时，虚假新闻司空见惯，“情感”与“想象”成为信息的内核，“事实”与“真相”被逐渐隐蔽、消解。如果不能得到及时控制，将导致传媒秩序被破坏，政府和媒体的公信力降低，进而带来社会恐慌、舆情肆意泛滥等极其严重的社会影响。为深入清理虚假新闻，营造风清气正的网络环境，中央网信办在全国范围内启动2024年度“清朗”专项行动。如何借助信息技术高效识别和治理虚假新闻已成为必须面对的时代课题。

目前，微博平台上存在的不实信息依然主要依赖用户的主动举报来处理。在接收到用户举报后，平台会派遣专员进行人工核实，最终在微博社区管理中心进行结果公示。这一过程不仅耗费大量的人力资源，还存在信息检测的滞后性。虚假新闻检测旨在利用机器学习等相关算法分析和处理新闻内容及其上下文，自动识别出可能存在虚假内容的新闻，能够有效降低人工审核的工作负荷，为网络信息质量监管提供重要技术保障。因此，学术界与工业界一直高度关注如何利用机器学习算法来自动地甄别虚假新闻。

当前的虚假新闻检测方法主要分为基于上下文[1-2]和基于内容两类方法[3-4]。基于上下文的方法通过分析信息传播路径、用户反馈等来追溯虚假新闻源头，但这种方法往往存在信息延迟，且有些用户甚至不会产生反馈，难以满足对虚假新闻早期发现的需求。基于内容的方法则专注于加强对新闻文本本身的语义理解，通过直接研判新闻内容中是否存在虚假新闻，以达到及早识别虚假新闻的目的。

早期研究侧重于复杂且昂贵的手工特征工程，结合传统的机器学习方法对真假信息进行分类。在过去的十余年中，具有强大表征学习能力的深度神经网络已被广泛应用于虚假新闻检测，通过自动提取高可区分性的特征，带来了性能的显著提升。

Transformer[5]模型的问世掀起了自然语言处理领域的狂潮，为语言模型的发展注入了关键的活力。基于Transformer的语言模型，通常称为预训练语言模型，如BERT系列模型[6]和GPT系列模型[7]，在各种自然语言处理应用中取得了显著进展。这些预训练语言模型通过在大规模数据集上的学习，能够发现并掌握隐含知识，进而获得更高质量的信息内容语义向量，使它们作为下游任务的特征表示器时，展现出远超传统深度学习模型的优越性能。然而，实践中也有研究指出，预训练小语言模型受限于参数规模与训练数据的不足，难以跨越信息边界，在处理复杂语义和整合多维度信息时存在局限性[8]。近年来，大语言模型凭借庞大的训练语料库、巨大的计算资源和海量的参数，极大提升了文本语义理解能力，并引发了传统的“预训练＋微调”范式向“预训练＋提示学习”范式的转变。然而，新闻的真实性通常是以丰富的证据和确切的事实作为支撑，仅依赖新闻本身的文本内容来判断其真伪，检测性能依然有限。

为进一步提升检测准确性，已有研究尝试引入外部证据，通过搜索引擎或专门的信息库，获取与新闻或帖子相关的网页、文章、标题和摘要等内容，为模型提供更多背景信息，帮助其更好地判断新闻的真实性[9]。但是，这种检索增强方式也面临一些挑战。首先，检索外部证据往往需要一定的计算资源和时间，导致成本较高。其次，由于外部证据的来源广泛且内容复杂，检索的结果容易返回包含噪声的文档或段落，这些噪声信息在某些情况下甚至可能误导模型的推理，导致模型产生错误的结论。简言之，检索增强方法在一定程度上提升了虚假新闻检测的准确性，但也存在不稳定和不可靠的风险[10]。随着大语言模型在语义理解能力方面变得愈发强大，为基于新闻内容的证据发现提供了新的研究思路。它们无需依赖外部检索方式，依据内在的庞大参数知识，能够综合多种维度的信息，直接生成证据文本。然而，大语言模型自身知识更新难、成本高，易引发幻觉问题，进而导致生成的证据中仍然可能存在一定的噪声。

鉴于此，本文提出了一种结合证据生成与语义融合的两阶段虚假新闻检测方法CollabDetection，主要贡献包括：

(1)秉承大小语言模型协同的思想，提出了一种基于证据生成和语义融合的虚假新闻检测方法CollabDetection，利用大语言模型生成潜在证据，微调预训练小语言模型，提取新闻文本内容和多维度证据特征，通过多模态因子分解双线性池化技术实现特征间的充分融合，提升信息聚合效率，充分发挥大小语言模型各自的优势。

(2)设计了一种自适应上下文构造策略，利用$k$近邻算法和最小描述长度原则，依据标注示例集为测试样本构建自适应的上下文语境数据。

(3)设计了七个证据评估维度，并模拟人类群体协作决策过程，提出了一种基于多大语言模型协商的证据生成方法mLLMNego。该方法使用不同的大语言模型分别充当生成器和判别器，采取自适应上下文构造策略，生成器生成多维度证据，判别器分析并评估证据的有效性，经过多轮交互，能够有效降低由幻觉问题引入的噪声，提升生成证据的稳定性和质量，从而避免了传统检索方法可能引入的噪声问题。

(4)在Twitter15、Weibo16和Pheme9三个真实数据集上进行了广泛的实验验证。结果表明，CollabDetection在大部分基准中均优于基于内容和基于特征融合的方法。其中，$F1$值在Twitter15数据集上提升了2.8%到46.6%，在Weibo16数据集上提升了0.3%到34.2%，在Pheme9数据集上提升了3.3%到59.3%。

## 2　相关工作

### 2.1　基于内容的虚假新闻检测方法

传统机器学习方法需要制定许多基于启发式规则和手工设计的特征，如统计特征、单词特征和句法特征，然后基于提取的特征训练有监督或无监督分类器对真假信息进行分类。Castillo等人[11]开创性地使用Twitter的文本统计特征如词数和表情符号，并采用决策树作为分类器。近年来，基于深度神经网络的方法在虚假新闻检测领域逐渐崭露头角，其优势在于能够自动从海量原始数据中提取高度区分的特征，无需烦琐的手工特征设计。例如，刘楠等人[4]研究事件级别的虚假新闻检测方法，结合图卷积网络GIN和Word2Vec模型通过事件传播图增强策略和对比学习来提高模型鲁棒性和泛化能力。尽管深度神经网络在处理局部或短期信息时表现优异，但在捕捉语义中的长期依赖关系和全局上下文时，往往存在局限性，对检测性能产生明显的制约。

针对中文社交媒体平台(如微博)的独特语言特性和传播模式，研究者们提出了专门的虚假新闻检测方法。具体而言，Nan等人[12]构建了一个面向中

文社交媒体的多领域虚假新闻数据集Weibo21，涵盖政治、军事、教育等九个领域。基于该数据集提出了MDFEND方法，通过领域门控机制动态聚合多专家网络的特征表示，解决了跨领域检测中由于词汇使用差异引起的数据分布偏移问题。针对微博平台的实时检测需求，黄学坚等人[13]提出了一种融合用户特征和内容特征的模型，该模型通过分析用户历史微博中的情感倾向，同时计算用户对特定领域的专业度，并引入表情符号、URL等非文本特征，显著提升了虚假新闻的早期识别能力。

在过去几年中，以BERT和GPT为代表的预训练语言模型因其强大的上下文语义感知能力为虚假新闻检测带来了新的见解。MetaAdapt[14]首先使用RoBERTa模型对输入进行编码，然后使用元学习技术快速迁移到新领域进行少样本虚假新闻检测，从而提高了模型跨不同领域的泛化能力。CRFB[15]是一种对比学习模型，随机屏蔽帖子中的词语生成噪音文本，使用BERT编码，以[CLS]Source[SEP]Comment[SEP]格式连接源帖和评论，并使用最终[CLS]标记的隐藏状态作为节点表示。虽然“预训练+微调”的范式在一定程度上可以捕捉上下文信息，但仍然受到模型容量有限的挑战。此外，社交媒体平台如推特和微博上的用户生成内容通常是稀疏、嘈杂且模糊的文本，预训练小语言模型在处理复杂语义结构和整合多维度信息时表现不足。

### 2.2 基于大语言模型的虚假新闻检测方法

近两年，大语言模型在各种自然语言处理任务中展现出优越的零样本和少样本学习能力，为虚假新闻检测开辟了新的可能性。例如，GenFEND[16]框架利用GPT-3.5模型模拟不同用户角色生成多样化评论，并按人口统计属性划分为多个子群体，通过视角内聚合与视角间融合生成综合用户反馈特征，结合新闻内容特征进行新闻真实性判别。柯婧等人[17]利用GPT-3.5-turbo模型对新闻内容的主干事件、细粒度次要事件和隐含信息进行层级推理总结，通过谷歌检索获取外部知识，最后选取LLaMA2-13B模型进行真实性判别，从而提升模型的检测性能和可解释性。尽管基于大语言模型的方法在虚假新闻检测中取得了一定的研究进展，但其局限性仍然显著。一方面，单大语言模型的单轮决策难以有效应对推理密集型和知识密集型任务。另一方面，这类方法对上下文选择的稳定性高度敏感，检测性能易受示例选择质量的影响[18]。

与此同时，研究者开始探索通用大语言模型和任务小语言模型在虚假新闻检测任务中的协同作用。大语言模型以其丰富的知识储备和强大的生成能力，弥补了小语言模型在低资源环境下的不足，而小语言模型则通过任务特定的微调，提供高效且精确的特征表示能力。L-Defense[19]框架首先利用RoBERTa模型从原始报道中提取支持和反对新闻声明的竞争性证据；然后，通过LLaMA2-7B模型分别对证据集合进行推理和总结，生成针对两种可能真实性的解释性文本；最后，结合由RoBERTa模型驱动的防御推断模块，对竞争性解释的质量进行建模和比较，从而判定新闻声明的真实性。Hu等人[20]设计了一种基于适应性推理指导的虚假新闻检测框架ARG，通过GPT-3.5模型生成多视角推理文本，再结合微调的BERT模型进行任务特定学习，对推理文本的质量进行评估和选择。文章中指出大语言模型能够从多个视角挖掘出合理的线索，但其在基于证据进行直接推理时的表现较差。

### 2.3 基于证据增强的虚假新闻检测方法

在现有虚假新闻检测研究中，大部分方法依赖新闻文本内容进行监督学习。这些方法通常通过分析新闻标题或社交媒体发布的文本来评估真实性。然而，忽视了可以进一步支撑判断的相关证据信息。因此，研究者引入证据增强的理念，通过不同的方式增强模型的判断能力。

许多方法依赖于从外部来源（如Wikipedia、社交媒体、新闻网站等）检索与新闻相关的证据。Popat等人[21]在2018年提出的DeClarE模型，是首个基于证据增强的深度学习模型，直接从网页中检索外部文章作为证据，利用双向LSTM建模语言特征，同时结合注意力机制聚焦于最相关的内容片段进行真实性判断。钟将等人[9]在此基础上创新地使用外部网页的标题和摘要作为证据。Liao等人[22]提出的MUSER框架模拟人类验证过程，结合多步证据检索，逐步增强证据质量，从而提升检测效果。GET模型[23]采用图结构建模，通过转化网络报道和发布者信息为图结构，捕捉长距离语义依赖，并通过动态移除冗余节点，聚焦于最相关的证据信息。尽管这些基于证据增强的虚假新闻检测方法已取得显著进展，但它们仍面临一些共同的挑战。首先，证据的质量和相关性至关重要，若证据不准确或与新闻内容不相关，可能会引入大量噪声，从而影响模型的判断；其次，许多证据检索和推理过程涉及复杂的计算和多轮检索，通常需要大量计算资源，可能导致响应时间过长。

受这些研究的启发，本文提出了基于证据生成和语义融合的虚假新闻检测方法CollabDetection，利用大语言模型生成多维度证据，并结合微调后的RoBERTa模型，对证据和新闻文本进行特征表示与融合，在复杂的虚假新闻检测任务中提供更为全面和细致的判断。

## 3　基于多大语言模型协商的多维度证据生成方法

本文提出了一种基于多大语言模型多轮协商的多维度证据生成方法mLLMNego，该方法由三个核心部分组成：生成器模块、鉴别器模块和协商机制。图1展示了mLLMNego的研究框架图，生成器和判别器分别使用不同的大语言模型。在证据生成阶段，首先根据已有研究提出七个证据评估维度，然后设计自适应上下文构造策略，为测试样本提供高相关性的上下文语境支持，从而提升生成器和判别器在证据生成过程中的稳定性。最后建立多大语言模型协商机制，具体包括三个轮次：第一轮，生成器基于自适应上下文提示生成初步证据文本；第二轮，鉴别器对生成器生成的候选证据进行评估，并提出修正建议，以优化生成器的下一轮推理方向；第三轮，生成器根据鉴别器反馈，修正之前的缺陷并进一步优化生成内容，最终输出一致性的推理结果。以下对各部分功能及其在多维度证据生成中的作用进行详细描述。

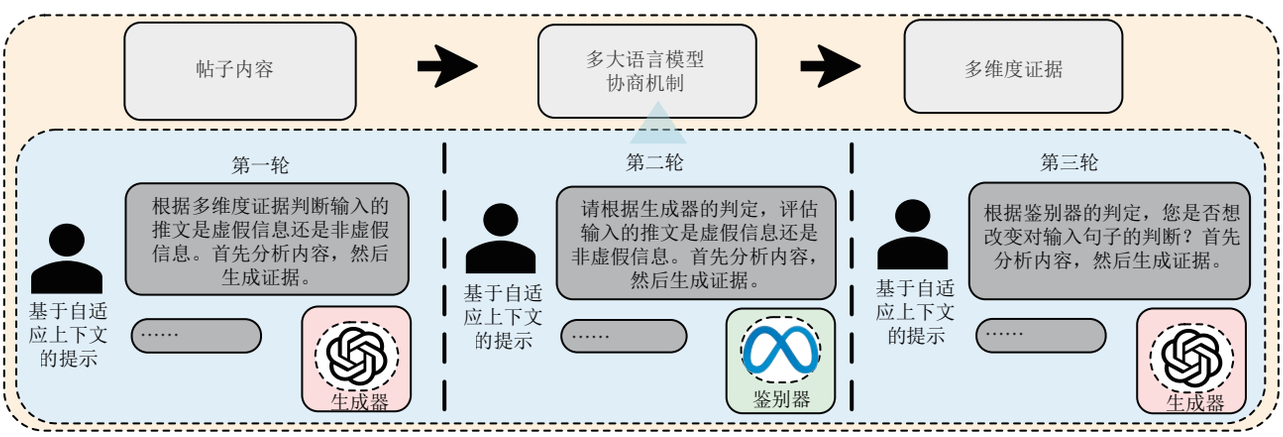


图1　基于生成-对抗思想和多轮协商机制的多维度证据生成框架

### 3.1　多维度证据定义

虚假新闻检测不仅仅是对内容真实性的简单判断过程，特别是在社交平台上，信息传播迅速，容易受到歪曲和误导。因此，需要从多个维度进行全面分析。根据相关研究，本文提出以下七个证据评估维度。

（1）信息准确性和完整性：信息的准确性和完整性直接影响其可信度。虚假新闻常通过曲解事实或选择性地呈现误导受众。正如Hu等人[20]所提到的，虚假新闻的创作者可能会操控新闻的任何部分，运用多种写作策略。此维度评估信息是否经过正确的解读并完整呈现了所有相关事实。

（2）信息情境关联性：信息情境关联性指的是信息是否适应其特定的社会、文化或政治背景。错误的情境关联可能导致公众误解并扩大虚假新闻的影响力。此维度评估信息是否正确地与特定的背景和情境相匹配。例如，某地五年前洪灾的救援报道未标注时间，被重新转发后，公众误认为该地区近期再遭灾害，导致恐慌性物资抢购。

（3）信息来源的可靠性：虚假新闻通常通过不可靠或未经验证的渠道传播，尤其是在社交平台上，信息的来源可能隐藏在大量的匿名账号中，难以追溯与验证。Min等人[2]的研究表明，社交媒体上的虚假新闻常通过恶意账号或虚假账号传播，这些账号缺乏清晰的社会背景和认证信息。此维度的核心在于评估信息来源是否经过认证，是否来自有公信力的媒体或专家机构。

（4）情感倾向：虚假新闻会利用情绪化的语言来操控受众的情感，促使他们产生极端反应。根据Zhang等人[24]的研究发现，信息发布者通过使用激烈的情感词汇，如“愤怒”、“恐惧”、“无助”等，引发公众的共鸣或恐慌，从而影响其判断。

（5）话题标签：在微博等社交平台中，话题标签作为信息传播的重要工具，有时也被滥用来扩大虚假新闻的影响力。虚假新闻发布者可能故意使用与热点事件相关的标签（如“#新能源车隐患”、“#食品

添加剂揭秘”），误导受众认为信息与真实事件相关。

（6）信息传播意图：传播信息的动机和意图是虚假新闻识别中的另一个关键线索。如Zhou等人[25]所述，虚假新闻传播者的意图可以分为有意传播与无意传播两种情形，前者通常具有恶意，而后者则可能由于被误导或缺乏信息准确性而无意间传播虚假内容。通过分析信息的传播动机，能够判断发布者是否有意图引导受众做出特定的反应。

（7）信息主题：根据Hu等人[26]的发现，虚假新闻检测与信息的主题密切相关，健康领域的新闻往往更容易被判定为虚假，而经济领域的新闻则更可能被认为是可信的。这些领域本身具有较高的社会关注度，能够迅速引发讨论。通过综合七个维度的分析，可以更全面、准确地评估新闻的真实性及潜在风险。多维度证据分析不仅帮助识别信息中的虚假成分，还能揭示信息传播中的操控手段，为检测提供科学依据。

### 3.2 自适应上下文构造策略

如图2所示，本文设计了一种样本级的自适应上下文构造策略，通过动态选择并排序上下文示例，为测试样本提供有序、高质量的上下文支持，从而提升证据推理的准确性与可靠性。首先，采用基于k近邻算法的搜索策略，从训练集中挑选出与待测试样本语义最为接近的$N$个示例。随后，将这些示例随机组合成多组候选上下文集合，并基于最小描述长度原则[27]评估每组示例的质量，最终选择描述长度最小的示例集合，作为大语言模型推理的上下文语境。

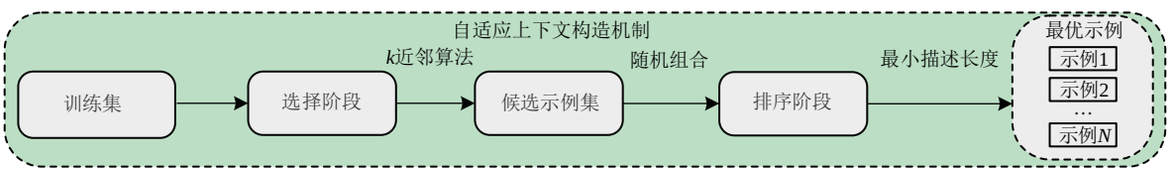


图2 自适应上下文构造策略

具体来说，给定一个测试样本$(x,y)$，基于上下文集合$c$使用生成式语言模型$\mathcal{P}$生成目标$y$的概率可以表示为：

$$p(y|c,x)=P\big(V(y)|c,T(x)\big) \tag{1}$$

其中，$\mathcal{T}(\cdot)$表示用于包装输入的模板，具体形式为“＜标签＞：＜新闻文本＞。上下文集合$c=\mathcal{T}(x_1),\cdots,\mathcal{T}(x_k)$是由$k$个输入-输出示例连接而成的上下文字符串。为了适配分类任务，引入了转化器$\mathcal{V}(\cdot)$，将每个标签$y$映射到生成式语言模型$\mathcal{P}$的词汇表中。

自适应上下文示例构造策略的目标是从候选上下文集合$C$中找到最佳的上下文集合$c^*$，使其能够针对当前测试输入$x$得到正确的标签$y$，并最大化模型性能。由于穷举搜索所有可能的上下文组合在计算上不可行，通过分为“选择阶段”和“排序阶段”高效缩小搜索空间。

在选择阶段，使用k近邻搜索方法，从训练集中选择与测试样本$x$语义最相似的k个示例。具体来说，首先计算测试样本$x$的嵌入向量$\boldsymbol{q}=\mathrm{Embedding}(x)$，然后通过内积计算相似度：

$$\mathrm{sim}(\boldsymbol{q},\boldsymbol{d}_i)=\boldsymbol{q}\cdot\boldsymbol{d}_i \tag{2}$$

其中，$\boldsymbol{d}_i$是训练集中的第$i$个示例的嵌入向量。从训练集的嵌入向量集合$\boldsymbol{D}=\{\boldsymbol{d}_1,\boldsymbol{d}_2,\cdots,\boldsymbol{d}_M\}$中选择与$\boldsymbol{q}$最相似的$k$个邻居，作为候选上下文集合$C_{\text{candidate}}$：

$$C_{\text{candidate}}=\mathrm{Topk}(\boldsymbol{q},\boldsymbol{D},k) \tag{3}$$

其中，$\mathrm{Topk}(\boldsymbol{q},\boldsymbol{D},k)$表示从$\boldsymbol{D}$中选择与$\boldsymbol{q}$最相似的$k$个示例，这个过程是基于相似度$\mathrm{sim}(\boldsymbol{q},\boldsymbol{d}_i)$排序的。通过这一阶段，能够快速缩小搜索范围，减少后续计算开销。

在排序阶段，基于最小描述长度原则，从信息压缩的视角评估候选集合的质量，通过计算数据的编码长度与模型编码长度的总和，选择既能解释数据又不过于复杂的上下文示例组合。通过最小化描述长度筛选出最优的上下文集合$c^*$：

$$c^*=\arg\min_{c\in C} L_\theta(y|c,x)+L(\theta) \tag{4}$$

其中，$\theta$为模型参数，$L_\theta(y|c,x)$表示在给定上下文集合$c$和测试输入$x$的情况下，对标签$y$的编码长度；$L(\theta)$表示模型的编码长度，在排序阶段，由于模型参数$\theta$不更新，可以忽略。$L_\theta(y|c,x)$使用香农信息理论中的负对数概率进行计算：

$$L_\theta(y|c,x)=-\log_2 p(y|c,x) \tag{5}$$

其中，$p(y|c,x)$表示模型在上下文集合$c$和输入$x$下对标签$y$的预测概率。如果模型对标签$y$的预测概

率越大，则相应的编码长度越小，表明模型对该标签的预测更有信心，因而该上下文集合$c$的质量更高。在排序阶段无法直接计算$p(y|c,x)$，因此引入期望值作为近似替代：

$$L_\theta(y|c,x)\approx-\mathbb{E}_{q(y_i|Y)}\log_2 p(y_i|c,x) \tag{6}$$

其中，$q(y_i|Y)$表示可能标签集合$Y$中$y_i$的先验分布。通过这种近似，模型评估了所有可能标签的加权平均编码长度，权重由先验分布$q(y_i|Y)$决定。当使用模型预测概率$p(y_i|c,x)$估计$q(y_i|Y)$时，等价于计算模型预测分布的熵。因此，计算上下文集合$c$的MDL损失：

$$L_{\mathrm{MDL}}(y_i|c,x)=-\mathbb{E}_{p(y_i|c,x)}\log_2 p(y_i|c,x) \tag{7}$$

为了增加上下文的多样性并提升模型的鲁棒性，从候选上下文集合$C_{\mathrm{candidate}}=\{c_1,c_2,\cdots,c_k\}$中随机选择B组上下文，每组包含$a$个示例：

$$C_{\mathrm{group}}^{(b)}=\{c_{b_1},c_{b_2},\cdots,c_{b_a}\},b=1,2,...,B \tag{8}$$

其中，$C_{\mathrm{group}}^{(b)}$是第$b$次随机组合得到的上下文集合。所有随机组合得到的上下文集合的并集表示为：

$$C_{\mathrm{combined}}=\bigcup_{b=1}^{B}C_{\mathrm{group}}^{(b)} \tag{9}$$

从$C_{\mathrm{combined}}$中选择MDL损失最小的上下文集合$c^*$：

$$c^*=\underset{C_{\mathrm{group}}^{(b)}\in C_{\mathrm{combined}}}{\arg\min}\ L_{\mathrm{MDL}}(y_i|C_{\mathrm{group}}^{(b)},x) \tag{10}$$

最终，输出优化后的上下文集合作为结果，供后续模块使用。

### 3.3　生成器模块

依据七个证据评估维度，生成器采用自适应上下文构造策略，构建多维度证据提示模板，使用大语言模型生成多维度推理证据。如图3所示，生成器初始提示模板由任务描述、示例部分和测试输入三部分组成。

任务描述明确模型需要完成的任务类型及输出格式，包括：(1)说明目标(例如，判定输入内容是否为虚假新闻，并生成多维度证据)；(2)定义证据维度(例如，给出具体的证据维度以及相关定义)(3)提供回答结构(例如，首先给出判定结果，然后提供简洁的理由)。示例部分由自适应上下文构造策略生成的有序高质量上下文示例组成。为了避免增加人工标注成本，此部分利用数据集中原有的标签(如“虚假”或“真实”)和对应新闻文本，以“<标签>:<新闻文本>”的形式呈现，为大语言模型提供清晰、具体的参考与类比对象，从而提升推断当前测试样本的准确性与相关性。测试输入是来自测试数据集的待检测文本，紧随任务描述和示例部分。最后，将所有任务描述、示例和测试输入整合到一个完整的模板中，确保推理结果的全面性与证据生成的鲁棒性。

根据多维度证据判断输入的推文是虚假新闻还是真实新闻。您的回答应清楚地解释您的决定，首先分析内容，然后生成证据。请分别考虑以下证据维度：

1.信息的准确性和完整性：分析信息是否存在曲解、误导或选择性夸大，评估其准确性和完整性。例如：检查信息是否包含扭曲事实、断章取义、隐瞒真相等手段。

2.信息的情景关联性：分析信息的背景和内容，确保信息在正确的情景中使用，防止拼凑剪接无关联内容。例如：检查信息是否将过时内容重新包装传播，或错误关联图文。

3.信息来源的可靠性：验证信息来源的真实性，防止来源仿冒和虚构来源。例如：确认来源是否为真实的新闻事件当事方或权威主体，是否存在冒充现象。

4.情感倾向：评估信息的情绪色彩是否用于影响读者的感知。例如：检查信息是否使用情感语言来影响意见。

5.话题标签：评估Hashtags是否有传播虚假信息或与热点话题对齐的潜力。例如：检查Hashtags是否用于扩大虚假信息的传播范围。

6.信息传播意图：明确信息传播者的意图，区分心理意图、社会意图、政治意图和经济意图。例如：识别信息传播者是否为满足不当心理需求、制造社会冲突、破坏政治生态或牟取经济利益。

7.信息主题：根据信息的内容主题，进一步分析和分类。例如：分析信息是否涉及日常生活、科学知识、公众热点、公众人物、政策法律等14类主题。

请先回答“虚假新闻”或“真实新闻”，然后提供支持您决策的证据解释，解释应大约为120字。

示例如下：

1.<标签>:<新闻文本>

2. <标签>:<新闻文本>

……

测试输入: .....

图3　生成器初始提示模板

### 3.4　鉴别器模块

鉴别器负责分析评估生成器的结果，判断其是否准确，并提供合理的解释和反馈。它不仅仅是一个评估工具，更是生成器决策优化的重要环节。通过对生成器结果的反驳或支持，鉴别器帮助优化多轮协商流程中的推理路径，最终生成一致性更强、更可靠的判定结果。图4展示了鉴别器的提示模板，由四个部分组成：任务描述、示例、测试输入和生成器的回答。

任务描述明确鉴别器的任务目标，例如“请根据生成器的判断，评估输入的推文是否为虚假新闻或真实新闻”。示例和测试输入与生成器的提示保持一致。在连接任务描述、示例集合和测试输入和生成器的回答后，将构建的提示输入鉴别器。鉴别器的回复包含两个关键内容：态度(例如，是/否)表示鉴别器是否同意生成器的判定；详细说明鉴别器为何支持/反驳生成器决策的理由。通过对多维度证据的详细评估，鉴别器不仅能够提供对生成器判定

请根据生成器的判定，评估输入的推文是虚假新闻还是真实新闻。您的回答应清楚地解释您的决定，首先分析内容，然后生成证据。请分别考虑以下证据维度：

1.信息的准确性和完整性：分析信息是否存在曲解、误导或选择性夸大，评估其准确性和完整性。例如：检查信息是否包含扭曲事实、断章取义、隐瞒真相等手段。

……

7.信息主题：根据信息的内容主题，进一步分析和分类。例如：分析信息是否涉及日常生活、科学知识、公众热点、公众人物、政策法律等14类主题。

请先回答“是或“否”，然后提供支持您决策的证据解释，解释应大约为120字。

示例如下：

1.<标签>:<新闻文本>

2.<标签>:<新闻文本>

……

测试输入：.....

生成器回答：.....

图4 鉴别器提示模板

的校验，还能通过结构化和清晰的反馈帮助生成器优化推理路径，提升证据生成的全面性、准确性。

### 3.5 协商机制

在多大语言模型多轮协商机制中，当鉴别器提出质疑或反驳生成器的判断时，生成器会进一步审视其初步决策，并决定是否根据鉴别器的反馈调整自己的判断。该机制模拟人类决策中多方协商的过程，通过多轮交互生成高质量的多维度证据文本。

图5展示了生成器在协商阶段中的提示模板，该模板由任务描述、示例、测试输入和鉴别器的回答四个部分组成。生成器的回复包括：态度（例如，是/否）表示生成器是否根据鉴别器的阐述改变初始判定；生成器选择坚持或调整判断的证据解释。

根据鉴别器的判定，您是否想改变对输入推文是虚假新闻的判断？您的回答应清楚地解释您的决定，首先分析内容，然后生成证据。请分别考虑以下证据维度：

1.信息的准确性和完整性：分析信息是否存在曲解、误导或选择性夸大，评估其准确性和完整性。例如：检查信息是否包含扭曲事实、断章取义、隐瞒真相等手段。

……

7.信息主题：根据信息的内容主题，进一步分析和分类。例如：分析信息是否涉及日常生活、科学知识、公众热点、公众人物、政策法律等14类主题。

请先回答“是”或“否”，然后提供支持您决策的证据解释，解释应大约为120字。

示例如下：

1. <标签>:<新闻文本>

2. <标签>:<新闻文本>

……

测试输入：.....

鉴别器回答：.....

图5 生成器协商提示模板

最终，协商结果的处理方式有三种：（1）当生成器和鉴别器的协商结果一致时，直接选择生成器的回答作为最终决策，并将其作为多维度证据文本使用。（2）当双方存在分歧时，生成器可能依据鉴别器的反馈调整判断，最终选择达成共识的协商结果作为决策依据。（3）如果多轮协商后仍无法达成一致，则可能出现结果“无法确定”的样本。当准确率计算中包含大量“无法确定”样本时，可能会出现两个问题：第一，当进行准确率计算时，这些“无法确定”的样本无法直接与真实标签进行对比，可能导致准确率计算结果的不完整或者不公平，尤其是在与基线模型比较时；第二，将“无法确定”样本从计算中剔除，可能导致评估结果失真，从而低估或高估某些模型的表现。为避免引入额外的不确定性，例如使用第三方模型进行最终判定，本文决定将所有“无法确定”样本视为生成器的初步判断结果，并将其纳入准确率计算，以确保公平比较。

为了更加清晰地呈现上述协商机制的执行流程，图6展示了生成器与鉴别器在多轮协商中的算法流程图。在输入阶段，系统接收测试样本，生成器基于样本生成初步证据及判断。在协商交互阶段，鉴别器对证据进行评估并提供反馈。如果双方判断一致，则直接输出生成器的回答；若存在分歧，生成器根据鉴别器的反馈决定是否调整证据。如果经过最大轮次的协商仍未达成一致，则输出生成器的初步判断作为最终结果。

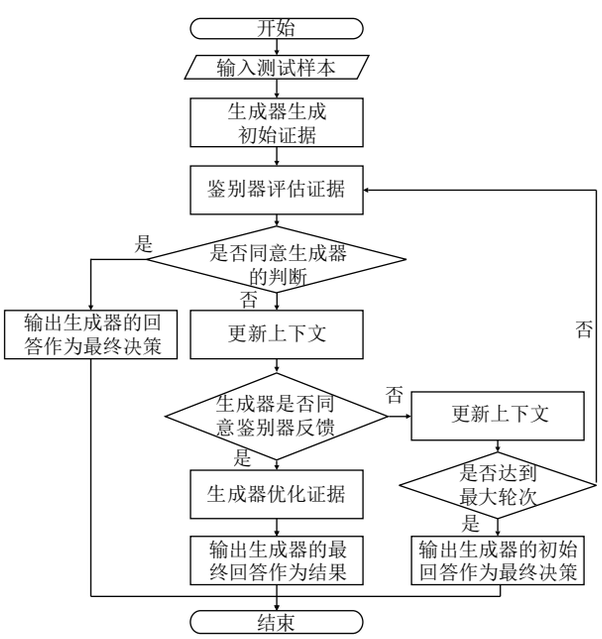


图6 多轮协商流程图

总之，多轮协商机制是生成高质量多维度证据的重要保障，不仅有效规避了传统检索方法引入的外部噪声，还在一定程度上抑制了大语言模型生成过程中的幻觉现象所带来的语义偏差。

## 4　虚假新闻检测方法

虚假新闻检测任务通常可以被定义为一个二分类问题。给定来自社交媒体的一组文本帖子，记为集合$X=\{x_1,x_2,\ldots,x_N\}$，目标是预测每个帖子的分类类别$Y=\{0,1\}$，其中$x_i$表示第$i$条具有一定字符长度的帖子，$N$代表文档数量，0表示真实新闻，1表示虚假新闻。通过模型训练，使用$f:X\rightarrow Y$将每个帖子映射到其对应的分类类别。

本文秉承“通用大语言模型＋任务小语言模型”协同思想，充分发挥大语言模型在广泛知识背景和多维度信息整合方面的优势，同时利用小语言模型在特定任务中的高效学习和精准语义表示能力，解决传统基于检索的证据增强方法的局限性。如图7所示，本文所提基于证据生成和语义融合的虚假新闻检测方法CollabDetection包括四个模块：(1)内容特征表征模块，微调预训练小语言模型(本文以RoBERTa模型为例)，提取新闻内容特征；(2)多维度证据特征表征模块，使用基于多大语言模型协商的多维度证据生成方法mLLMNego生成多维度证据文本，并将其输入小语言模型进行特征表示，获得多维度证据特征；(3)MFB融合模块，采用多模态因子分解双线性池化技术聚合内容特征和多维度证据特征；(4)虚假新闻检测器，将融合后的特征输入至多层感知机，经过处理后，使用softmax进行分类。接下来，本文将对这些模块逐一进行深入描述。

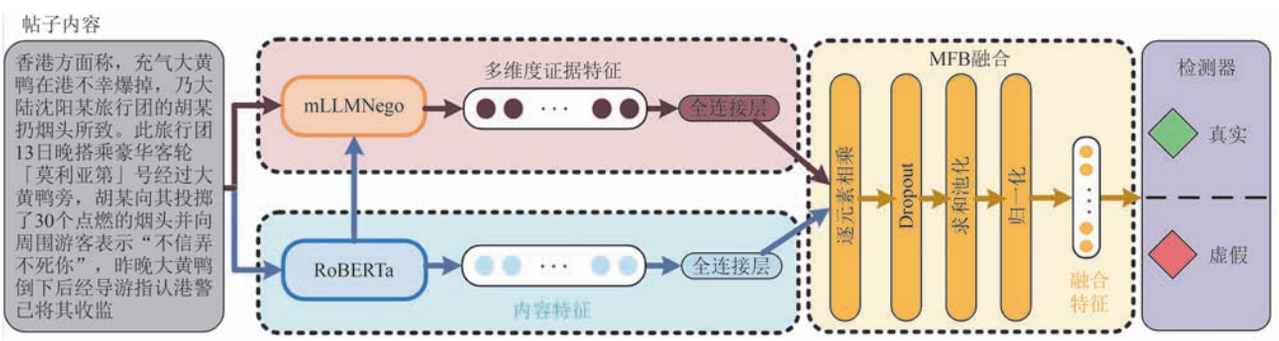


图7　融合多特征语义的虚假新闻检测框架

### 4.1　内容特征表征模块

本文选择RoBERTa模型作为编码器，并在虚假新闻数据集上对其进行微调。对于给定的输入文本序列$x_i=\{x_{i1},\cdots,x_{in}\}$，首先对其进行标记化处理。添加分类标记[CLS]和分隔符标记[SEP]后，得到标记化列表$V=\{[\mathrm{CLS}],v_1,\cdots,v_n,[\mathrm{SEP}]\}$。这里，$n$是帖子$x_i$中的单词数量。然后将这些标记输入模型，生成词嵌入$E=\{E[\mathrm{CLS}],E_1,\cdots,E_n,E[\mathrm{SEP}]\}$。通过多层Transformer编码器，RoBERTa产生词上下文输出$T=\{T[\mathrm{CLS}],T_1,\cdots,T_n,T[\mathrm{SEP}]\}$。最终，RoBERTa模型最后一层中与[CLS]标记对应的语义向量$T[\mathrm{CLS}]$，即维度为d的向量，被认为是帖子$x_i$的内容特征，将其简化为标记$\boldsymbol{Q}_C$。

$$\boldsymbol{Q}_C=T[\mathrm{CLS}]\in\mathbb{R}^d \tag{11}$$

### 4.2　多维度证据特征表征模块

本模块在内容特征表征模块的基础上，利用微调后的RoBERTa模型对mLLMNego生成的多维度证据进行特征提取。首先，通过自适应上下文示例构造策略从训练集中得到相关示例集合。随后，GPT-4o负责生成初步证据，Llama 3则作为鉴别器，通过多轮协商优化证据，mLLMNego得到涵盖七个维度的高质量证据文本。

对应于帖子文本$x_i$生成的证据文本$h_i=\{h_{i1},\cdots,h_{im}\}$输入到微调后的RoBERTa模型进行特征提取。与内容特征表征模块类似，RoBERTa将证据文本编码为上下文向量输出$T_{\mathrm{evidence}}=\{T_{\mathrm{evidence}}[\mathrm{CLS}],T_1,\cdots,T_m,T_{\mathrm{evidence}}[\mathrm{SEP}]\}$。其中，$T_{\mathrm{evidence}}[\mathrm{CLS}]$作为与帖子$x_i$对应的多维度证据特征，简化为标记$\boldsymbol{Q}_e$：

$$\boldsymbol{Q}_e=T_{\mathrm{evidence}}[\mathrm{CLS}]\in\mathbb{R}^{\mathrm{d}} \tag{12}$$

### 4.3　特征融合模块

特征融合模块的目标是将内容特征$\boldsymbol{Q}_C\in\mathbb{R}^{\mathrm{d}}$和多维度证据特征$\boldsymbol{Q}_e\in\mathbb{R}^d$作为输入，输出融合特征$F\in\mathbb{R}^o$。然而，直接拼接多种语义特征无法捕捉到特征间的语义关联，导致预测性能不佳。为了解决这一问题，本研究采用多模态因子分解双线性池化(MFB)技术[28]进行多特征语义融合。该技术使用双线性操作捕捉高阶特征关系，并通过低秩分解优

化计算效率，在视觉问答领域取得不错的效果。为了将该技术应用于多种文本特征的融合，本文将通道数设置为1。

具体而言，将帖子内容特征$\boldsymbol{Q}_C$和多维度证据特征$\boldsymbol{Q}_e$分别映射到双线性空间，通过投影矩阵$\boldsymbol{W}=[\boldsymbol{W}_1,\boldsymbol{W}_2,\cdots,\boldsymbol{W}_o]\in\mathbb{R}^{d\times d\times o}$，计算双线性交互特征。将双线性操作的输出特征表示为：

$$F_i = \boldsymbol{Q}_C^{\mathrm{T}}\boldsymbol{W}_j\boldsymbol{Q}_e \tag{13}$$

其中，$\boldsymbol{W}_j\in\mathbb{R}^{d\times d}$是投影矩阵，$\mathrm{F}_i$是第$i$个双线性层的输出。需要注意的是，由于每个双线性池化层引入大量参数，导致计算复杂度较高。为了降低计算复杂度，投影矩阵$\boldsymbol{W}_j$被分解为两个低秩矩阵$\boldsymbol{P}_i\in\mathbb{R}^{\mathrm{d}\times\mathrm{b}}$和$\boldsymbol{Z}_i\in\mathbb{R}^{\mathrm{d}\times\mathrm{b}}$，公式调整为：

$$F_i=\boldsymbol{Q}_C^{\mathrm{T}}\boldsymbol{P}_j\boldsymbol{Z}_j^{\mathrm{T}}\boldsymbol{Q}_e=1^{\mathrm{T}}\left(\boldsymbol{P}_j^{\mathrm{T}}\boldsymbol{Q}_C\circ\boldsymbol{Z}_j^{\mathrm{T}}\boldsymbol{Q}_e\right) \tag{14}$$

其中，$\circ$表示哈达玛积，即两个向量的元素逐位相乘。$1\in\mathbb{R}^{\mathrm{b}}$是全为1的向量，用于对逐元素乘积结果进行求和。

为了进一步获得最终的融合特征$F\in\mathbb{R}^{\mathrm{o}}$，需要学习两个张量$\boldsymbol{P}=[\boldsymbol{P}_1,\cdots,\boldsymbol{P}_{\mathrm{o}}]\in\mathbb{R}^{\mathrm{d}\times\mathrm{b}\times\mathrm{o}}$和$\boldsymbol{Z}=[\boldsymbol{Z}_1,\cdots,\boldsymbol{Z}_{\mathrm{o}}]\in\mathbb{R}^{\mathrm{d}\times\mathrm{b}\times\mathrm{o}}$。融合特征通过将$P$和$Z$重塑为矩阵$\tilde{\boldsymbol{P}}\in\mathbb{R}^{\mathrm{d}\times\mathrm{bo}}$和$\tilde{Z}\in\mathbb{R}^{\mathrm{d}\times\mathrm{bo}}$，然后进行求和和池化，使用大小为$r$的非重叠窗口聚合特征元素。这一过程可以形式化表示为：

$$\begin{aligned}F&=\mathrm{MFB}(\boldsymbol{Q}_C,\boldsymbol{Q}_e)\\&=\mathrm{SumPooling}(\tilde{\mathbf{P}}^{\mathrm{T}}\boldsymbol{Q}_C\circ\tilde{\mathbf{Z}}^{\mathrm{T}}\boldsymbol{Q}_e,r)\end{aligned} \tag{15}$$

最后，为了确保融合特征的稳定性和表达能力，采用以下策略：(1)引入dropout机制避免模型过拟合。(2)引入幂归一化($F\leftarrow\mathrm{sign}(F)\sqrt{|F|}$)和$L_2$归一化$\left(F\leftarrow\frac{F}{\|F\|}\right)$消除特征之间的尺度差异。

### 4.4 预测模块

在特征融合模块生成高质量的融合语义特征F后，基于MLP构建虚假新闻检测器。首先对融合特征进行线性变换和非线性激活操作，随后通过softmax映射为概率分布。预测公式为：

$$\hat{y}=\mathrm{softmax}\left(\mathrm{MLP}(F)\right) \tag{16}$$

其中，$\hat{y}$表示虚假新闻检测的预测标签。

为了优化模型性能，训练过程采用交叉熵损失函数，并加入$\mathrm{L}_2$正则化项以防止模型过拟合。目标损失函数定义为：

$$L=-\frac{1}{N}\sum_{i=1}^{\mathrm{N}}\left(y_i\log\hat{y}_i+\left(1-y_i\right)\log\left(1-\hat{y}_i\right)\right)+\lambda\mathrm{L}_2(\Theta) \tag{17}$$

其中，N是样本数，$y_i\in\{0,1\}$是第$i$个样本的实际标签，$\hat{y}_i$是第$i$个样本的预测标签，λ是$\mathrm{L}_2$正则化系数，$\mathrm{L}_2(\Theta)$是用于正则化模型参数$\Theta$的$\mathrm{L}_2$范数。

## 5 实验与分析

### 5.1 数据集

为了验证本文所提方法CollabDetection的有效性，在Twitter15[29]、Weibo16[3]、Pheme[30]三个公开中英文数据集上进行了大量实验，数据集的统计数据如表1所示。

将每个数据集按照8:2的比例划分为训练集和测试集，在证据生成阶段，训练集用于生成自适应上下文示例，测试集用于验证模型性能。

### 5.2 实验设置

本文所有实验均在一台配备有两块NVIDIA Tesla-V100 32 GB GPU的服务器上运行，算法实现基于PyTorch框架。

在证据生成阶段，对于自适应上下文构造策略的实验，使用Hugging Face平台的“sentence-transformers/all-mpnet-base-v2”模型提取测试数据的嵌入向量，并通过k近邻检索筛选30个候选示例。随机组合生成10组上下文，每组包含3个示例，利用“openai-community/gpt2-xl”模型计算最小描述长度得分评估上下文质量。为了评估所提多维度证据生成方法mLLMNego的公平性与有效性，分别选择了OpenAI的闭源模型(GPT-3.5和GPT-4o)、Meta AI的开源模型(Llama 2和Llama 3)以及阿里云的开源模型(Qwen 2.5)作为基线。其中，OpenAI的模型通过调用API接口“gpt-3.5-turbo”和“gpt4o”版本；Meta AI的模型采用Hugging Face平台提供的“meta-llama/Llama-2-7b-chat-hf”和“meta-llama/Meta-Llama-3-8B-Instruct”版本；阿里云的模型采用“Qwen/Qwen2.5-7B-Instruct”。

在语义融合阶段，中文数据集Weibo16使用“hfl/chinese-roberta-wwm-ext”，“google-bert/bert-base-chinese”，和“hfl/chinese-macbert-base”模型；英文数据集Twitter15、Pheme9使用“FacebookAI/roberta-base”，“google-bert/bert-base-uncased”，和“albert/albert-base-v2”模型。内容特征表征模块采用BertForSequenceClassification架构，设置最大序列长度为512，输入向量的维度为768。训练过程中，使用AdamW优化器，学习率设置为2e-5，并在

500个预热步骤后进行线性衰减。批量大小设置为4，训练周期为5，丢弃率设为0.1。特征融合模块的窗口大小$r$设为1，系数$\lambda$设为0.01。预测模块使用了交叉熵损失函数和Adam优化器，其中初始学习率为1e-4，训练周期数为200，并应用了早停机制以避免过拟合。

最后，本文采用准确率（Accuracy）、精确率（Precision）、召回率（Recall）以及F1值（F1-score），对所提方法性能进行了全面评估。

**表1　数据集分布统计**

| | 信息类别 | Twitter15 | Weibo16 | Pheme9 |
|---|---|---|---|---|
| 训练集 | 虚假新闻 | 296 | 1841 | 1922 |
| | 真实新闻 | 297 | 1889 | 3218 |
| | 总计 | 593 | 3730 | 5140 |
| 测试集 | 虚假新闻 | 74 | 471 | 480 |
| | 真实新闻 | 75 | 462 | 805 |
| | 总计 | 149 | 933 | 1285 |
| 信息总数 | 虚假新闻 | 370 | 2312 | 2402 |
| | 真实新闻 | 372 | 2351 | 4023 |
| | 总计 | 742 | 4663 | 6425 |

### 5.3　对比实验

#### 5.3.1　对比方法

为了验证CollabDetection的有效性，本文挑选了五类基线模型：传统深度神经网络模型（TextCNN、Bi-GRU、Transformer）、微调的预训练小语言模型（BERT、RoBERTa）、基于提示的大语言模型（GPT-3.5、GPT-4o、Llama 2、Llama 3、Qwen 2.5）、多特征语义融合模型（RDEA、DualEmo、CausalRD、CCFD、SBAG、GACL、IRDNet）和基于检索的证据增强模型（Search）进行了一系列对比实验。其中，为了保证实验结果的公平性，所有开源大语言模型的参数规模控制在7B至8B范围，以排除参数量级差异对模型性能评估的干扰。五类基线模型具体如下。

（1）TextCNN[31]：该方法利用卷积神经网络（CNN）提取局部特征，并通过全连接层和softmax操作实现虚假新闻分类。

（2）Bi-GRU[3]：作为LSTM的一种变体，Bi-GRU分别从前向和后向处理输入序列。分类器接收来自两个方向的最后一层隐藏层的连接输出用于虚假新闻检测。

（3）Transformer[5]：一种仅使用注意力机制而非卷积神经网络和循环神经网络构建的深度神经网络模型，因此在建模长序列方面具有更强的能力。

（4）BERT[6]：该方法利用多层双向变压器学习帖子的语义表示，通过[CLS]标记进行最终分类。

（5）RoBERTa[32]：作为BERT的优化版本，RoBERTa通过更大的数据集、更长的训练时间以及移除BERT中的Next Sentence Prediction任务，进一步提升了模型的语义表示能力。

（6）GPT-3.5[33]：由OpenAI提供的训练遵循人类指令的大语言模型，具备强大的自然语言理解和生成能力，广泛应用于各种自然语言处理任务。

（7）GPT-4o：OpenAI推出的最新旗舰人工智能模型，具有增强的速度和性能，支持文本、语音和视觉等多模态处理能力。

（8）Llama 2[34]：由Meta AI发布的开源语言模型，采用优化后的Transformer架构，在大规模语料上进行预训练，支持对下游任务的快速适应。

（9）Llama 3[35]：Llama 2的升级版本，在参数规模和训练策略上进一步优化，具备更强的语言生成和推理能力。

（10）Qwen2.5：阿里云研发的开源大语言模型，经过多领域优化，支持中文和英文等多语言任务，具备较强的跨领域适应性。

（11）mLLMNego：本文提出的基于多大语言模型协商的多维度证据生成方法，通过自适应上下文构造策略和基于生成器-鉴别器的多轮对抗协商优化推理过程，生成多维度自然语言证据，实现更全面的真实性判定。

（12）RDEA[36]：设计了三种事件增强策略，结合图卷积网络和自监督对比学习，学习事件的图结构表示。模型通过特征拼接的方式，将源帖的文本特征与图卷积网络提取的事件图特征进行融合。

（13）DualEmo[24]：认为发布帖子的情感有助于判断帖子的真实性。使用双重情感特征作为HSA-BLSTM的增强，并同时考虑新闻内容和评论。

（14）CausalRD[37]：一种基于因果图的虚假新闻检测框架，有效结合了去偏用户行为和事件内容信息，采用图卷积网络对传播结构和用户行为进行建模，最终通过特征拼接的方式，融合用户偏好、文本特征和事件图结构。

（15）CCFD[38]：认为负样本未得到充分探索，因此需要使用课程学习自动选择和训练负样本。模型通过图卷积网络建模新闻的传播结构，并结合源帖特征与响应帖的中心向量，最终通过特征拼接融合这些信息，生成综合新闻表示。

(16)SBAG[39]:模型通过将基于图卷积的用户发布特征、基于图注意力的用户互动特征和基于卷积神经网络的文本特征进行拼接形成一个综合特征表示,输入全连接层进行虚假新闻检测。

(17)GACL[40]:一种对比学习模型,将图卷积网络提取的图结构特征、BERT编码的文本特征和对抗训练生成的特征进行拼接,能有效处理噪声和对抗样本。

(18)IRDNet[41]:采用多任务训练框架,利用BERTweet结合BiLSTM和胶囊网络提取多层次的语义特征,同时通过事件级别和意图级别的对比学习策略,融合语义特征与意图特征。

(19)Search:本文结合Google的Programmable Search Engine和Google Search API,将新闻文本作为查询关键词,整合前5个网页对应片段作为新闻内容的搜索证据文本,并使用RoBERTa编码得到搜索证据特征,最后与内容特征拼接进行分类。

5.3.2 结果分析

在本节中,本文将基于证据生成和语义融合的虚假新闻检测方法CollabDetection与19个基准方法进行了比较,结果如表2、表3和表4所示,本文提出的CollabDetection在三个数据集上达到了SOTA性能。其中,星号(*)表示这些结果是引用自原始论文的方法,短横线(-)表示对应指标未在论文中呈现。分析可得以下结论:

(1)Transformer的表现略逊于TextCNN和Bi-GRU,这一现象可能归因于Transformer模型的归纳偏差较低,需要更多的训练数据才能充分发挥优势。此外,大多数基于深度神经网络的方法在性能上仍无法与经过微调的预训练小语言模型竞争。

(2)大语言模型在零样本场景下的表现较弱,例如在Twitter15数据集上,Qwen2.5和Llama 2的F1值仅为0.500和0.651;而在Pheme9数据集上,LLaMA 3的F1值甚至下降至0.466。这是因为零样本提示依赖模型的内在推理能力,缺乏上下文示例的支持,导致复杂语境下的表现不足。此外,Pheme9数据集本身的结构也加剧了这一挑战。该数据集围绕9个具体事件构建,每个事件中既包含真实新闻,也包含虚假新闻。这种"同事件-异立场"的分布特性会使得模型更难区分真假信息。然而,GPT-4o在零样本场景下相较其他的大语言模型表现相对优异,这表明更高参数规模和优化训练策略的大模型在推理任务中具有更好的鲁棒性。

(3)mLLMNego大幅领先其他大语言模型,且缩小了与微调的预训练小语言模型的差距,说明本文所提七个证据评估维度的有效性,以及经过多大语言模型协商之后生成的多维度证据能够更全面地捕捉虚假新闻特征,从而显著提升虚假新闻检测能力。尤其在Weibo16数据集上,F1分数达到了0.927,与RoBERTa的0.931相媲美,远超DualEmo的0.844。

**表2 CollabDetection与其他基线在Twitter15数据集上的性能比较**

| 方法 | 准确率 | 精确率 | 召回率 | F1值 |
|---|---|---|---|---|
| TextCNN | 0.920 | 0.908 | 0.932 | 0.920 |
| Bi-GRU | 0.913 | 0.918 | 0.905 | 0.912 |
| Transformer | 0.893 | 0.867 | 0.932 | 0.900 |
| BERT | 0.879 | 0.912 | 0.838 | 0.873 |
| RoBERTa | 0.940 | 0.958 | 0.919 | 0.938 |
| GPT-3.5 | 0.685 | 0.634 | 0.865 | 0.731 |
| GPT-4o | 0.725 | 0.709 | 0.757 | 0.732 |
| Llama 2 7B | 0.503 | 0.500 | 0.932 | 0.651 |
| Llama 3 8B | 0.691 | 0.694 | 0.676 | 0.685 |
| Qwen2.5 7B | 0.611 | 0.690 | 0.392 | 0.500 |
| RDEA* | 0.855 | - | - | 0.880 |
| CausalRD* | 0.862 | - | - | 0.886 |
| CCFD* | 0.856 | - | - | 0.877 |
| SBAG | 0.917 | 0.928 | 0.917 | 0.921 |
| GACL* | 0.901 | - | - | 0.877 |
| IRDNet* | 0.917 | - | - | 0.917 |
| Search | 0.846 | 0.859 | 0.824 | 0.841 |
| mLLMNego(Ours) | 0.886 | 0.913 | 0.851 | 0.881 |
| CollabDetection(Ours) | **0.966** | **0.973** | **0.959** | **0.966** |

**表3 CollabDetection与其他基线在Weibo16数据集上的性能比较**

| 方法 | 准确率 | 精确率 | 召回率 | F1值 |
|---|---|---|---|---|
| TextCNN | 0.884 | 0.915 | 0.849 | 0.881 |
| Bi-GRU | 0.875 | 0.863 | 0.894 | 0.878 |
| Transformer | 0.842 | 0.838 | 0.854 | 0.845 |
| BERT | 0.928 | 0.923 | 0.936 | 0.929 |
| RoBERTa | 0.931 | 0.943 | 0.919 | 0.931 |
| GPT-3.5 | 0.791 | 0.777 | 0.822 | 0.799 |
| GPT-4o | 0.889 | 0.880 | 0.902 | 0.891 |
| Llama 2 7B | 0.603 | 0.570 | 0.875 | 0.690 |
| Llama 3 8B | 0.770 | 0.744 | 0.828 | 0.784 |
| Qwen2.5 7B | 0.711 | **0.963** | 0.444 | 0.608 |
| DualEmo | 0.844 | 0.844 | 0.844 | 0.844 |
| SBAG | 0.946 | 0.947 | 0.946 | 0.946 |
| Search | 0.933 | 0.942 | 0.924 | 0.933 |
| mLLMNego(Ours) | 0.925 | 0.907 | **0.949** | 0.927 |
| CollabDetection(Ours) | **0.950** | 0.957 | 0.943 | **0.950** |

（4）多特征语义融合模型的性能普遍较好，且特征融合策略对性能提升至关重要。实验结果显示，SBAG和本文提出的CollabDetection优于其他基线方法。相较CollabDetection，SBAG通过简单拼接用户发布特征、互动特征和文本特征实现融合。在Twitter15和Weibo16两个数据集上，CollabDetection的F1分数比SBAG分别提高了4.5%和0.4%，在Pheme9数据集上CollabDetection的F1分数比GACL提高了6.4%，验证了MFB融合策略能够更高效地聚合内容特征和多维度证据特征，捕捉复杂的特征交互，提升模型的整体性能。

（5）证据增强模型的性能很大程度上取决于提供的证据的质量。相较基线模型RoBERTa，基于检索的证据增强模型Search在Weibo16和Pheme9数据集上性能相当，但在Twitter15数据集上F1值降低了9.7%。然而，相较Search，本文所提CollabDetection在三个数据集上提升了1.7%到12.5%。这表明，传统检索方式提供的证据可能包含冗余、片面甚至与原始新闻无关的信息，这些“噪声证据”会对模型判断造成干扰，而本文所提多大语言模型协商机制能够有效降低证据生成过程中的噪声，使得证据的质量和稳定性更高。

（6）协同利用通用大语言模型和任务小语言模型，可以充分发挥大语言模型在广泛知识背景和多维度信息整合方面的优势，以及小语言模型在特定任务中的高效学习和精准语义表示能力。与其他基线方法相比，本文提出的CollabDetection方法在Twitter15、Weibo16和Pheme9的四个评价指标上都取得了优异的结果。在Twitter15数据集上$F1$值提升2.8%到46.6%，在Weibo16数据集上$F1$值提升0.3%到34.2%，在Pheme9数据集上$F1$值提升3.3%到59.3%。

### 5.4　自适应上下文构造策略的有效性分析

为了验证自适应上下文构造策略对虚假新闻检测性能的影响，本文在Weibo16和Twitter15两个数据集上分别评估了不同的示例选择方法，包括零样本、少样本（随机）以及少样本（自适应）。所有实验无需额外训练模型。在零样本场景下，大语言模型直接通过提示对测试样本进行真实性判定，无需提供示例上下文。在少样本场景中，大语言模型结合自适应上下文示例构造策略生成的有序高质量上下文进行推理。值得注意的是，为了避免增加人工标注成本，示例部分由数据集中原有的标签（如“Fake news”或“Real news”）和对应新闻文本组成，即以“<标签>:<新闻文本>”的形式呈现。实验结果如图8和图9所示。

**表4　CollabDetection与其他基线在Pheme9数据集上的性能比较**

| 方法 | 准确率 | 精确率 | 召回率 | F1值 |
|---|---|---|---|---|
| TextCNN | 0.838 | 0.800 | 0.756 | 0.777 |
| Bi-GRU | 0.817 | 0.741 | 0.785 | 0.762 |
| Transformer | 0.821 | 0.769 | 0.744 | 0.756 |
| BERT | 0.879 | 0.829 | 0.850 | 0.840 |
| RoBERTa | 0.884 | 0.837 | 0.856 | 0.847 |
| GPT-3.5 | 0.645 | 0.531 | 0.431 | 0.476 |
| GPT-4o | 0.636 | 0.518 | 0.367 | 0.429 |
| Llama 2 7B | 0.391 | 0.378 | **0.977** | 0.545 |
| Llama 3 8B | 0.581 | 0.444 | 0.490 | 0.466 |
| Qwen2.5 7B | 0.612 | 0.459 | 0.223 | 0.300 |
| GACL* | 0.850 | 0.836 | 0.826 | 0.829 |
| Search | 0.842 | 0.893 | 0.879 | 0.860 |
| mLLMNego(Ours) | 0.750 | 0.635 | 0.779 | 0.700 |
| CollabDetection(Ours) | **0.893** | **0.894** | 0.893 | **0.893** |

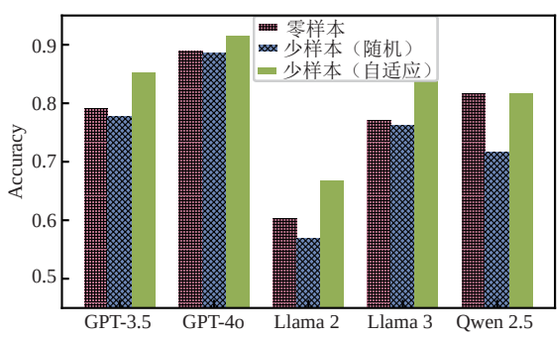


图8　不同示例选择方法在Weibo16数据集上的影响

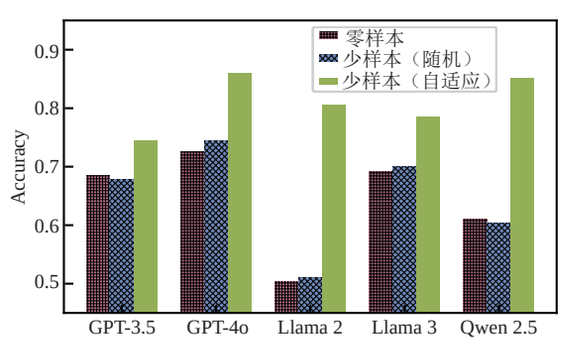


图9　不同示例选择方法在Twitter15数据集上的影响

实验结果显示，自适应策略在两个数据集上的表现明显优于零样本和随机选择方法。以Llama3模型为例，在Twitter15数据集上，采用自适应策略后，准确率从零样本基线的69.1%提升至78.5%，提升幅度达到9.4%；在Weibo16数据集上，自适应策略相较于随机选择方法，准确率提高了7.5%。

这足以表明，自适应上下文构造策略能够提供更相关的上下文示例，从而显著提高推理效果。

在Twitter15数据集上，随机示例选择方法的性能都低于零样本方法，在Weibo16数据集上，GPT-3.5和Qwen2.5也是如此。出现这一反常现象的原因可能是，一方面，示例构成中没有提供思维链形式的推理过程，另一方面，随机方式引入了不相关甚至矛盾的上下文示例，增大了大语言模型推理的难度。但采用本文所提的自适应上下文构建策略，大语言模型的推理准确性得到大幅提升。

此外，为评估自适应上下文策略中k近邻参数（k值）对模型性能的影响，本文基于生成器GPT-4o在Twitter15和Weibo16数据集上开展对比实验。实验中，k值分别设置为15、30和45，结果如图10所示。在Twitter15数据集上，$k$=15和$k$=30的模型准确率相近，而$k$=45的准确率略有下降。在Weibo16数据集上，$k$=30的准确率最高。综合两个数据集的实验结果，$k$=30在准确率与计算效率之间实现了较优平衡。

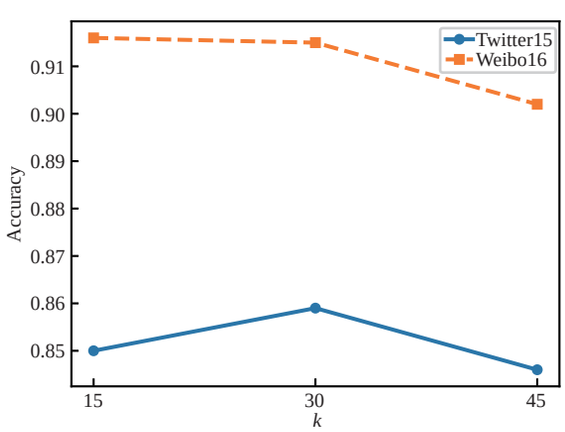


图10 不同k值在Twitter15和Weibo16数据集上的影响

### 5.5 多大语言模型协商机制实验分析

为了深入探讨本文提出的生成器主导推理、鉴别器辅助校验的多大语言模型协商机制的优越性，在Weibo16和Twitter15两个数据集上进行架构对比实验，评估不同大语言模型作为生成器（G）和鉴别器（D）时的性能表现。实验涵盖了无协商（大语言模型仅作为生成器）、自我协商（同一模型担任生成器和鉴别器）以及跨模型协商（不同模型担任生成器和鉴别器）三种架构配置，所有实验均基于自适应上下文构造机制。初步实验结果表明，GPT-4o作为生成器的表现优于其他模型，因此选择GPT-4o作为生成器，配合其他大语言模型作为鉴别器，进行跨模型协商的实验。

从表5中可以观察到。

（1）在所有实验中，GPT-4o作为生成器的表现始终优于其他模型，无论是在无协商、自我协商还是跨模型协商中。例如，在Twitter15和Weibo16数据集上，mLLMNego的$F1$值比Llama 3（G）+GPT-4o（D）的$F1$值分别高出4.6%和5.2%，展现出强大的生成能力和推理能力。

（2）具有鉴别器辅助的对抗协商架构在大多数情况下优于无对抗协商架构。例如，在Twitter15数据集上，GPT-4o作为生成器单独使用时的$F1$值为0.859，而与Llama 3作为鉴别器时，$F1$值提升至0.881。结果表明仅依赖生成器可能无法充分挖掘任务潜力，而鉴别器可以优化推理结果。然而，当同一模型既担任生成器又担任鉴别器时，通常会由于模型能力的重叠而限制整体表现。以GPT3.5（G）+GPT3.5（D）为例，F1值在Weibo16数据集上仅为0.786，低于无协商和跨模型协商架构，这表明生成器与鉴别器的角色应当加以区分，并由不同模型来承担。

**表5 不同的多大语言模型协商架构的性能比较**

| 方法 | Twitter15 | | Weibo16 | |
|---|---|---|---|---|
| | 准确率 | F1值 | 准确率 | F1值 |
| GPT-3.5 (G) | 0.745 | 0.729 | 0.852 | 0.852 |
| GPT-4o (G) | 0.859 | 0.859 | 0.915 | 0.916 |
| Llama 3 (G) | 0.785 | 0.758 | 0.837 | 0.841 |
| Qwen2.5 (G) | 0.852 | 0.845 | 0.817 | 0.789 |
| GPT-3.5 (G) + GPT-3.5 (D) | 0.758 | 0.731 | 0.805 | 0.786 |
| GPT-4o (G) + GPT-4o (D) | 0.872 | 0.869 | 0.922 | 0.924 |
| Llama 3 (G) + Llama 3 (D) | 0.792 | 0.780 | 0.841 | 0.850 |
| Qwen2.5 (G) + Qwen2.5 (D) | 0.779 | 0.752 | 0.840 | 0.822 |
| GPT-3.5 (G) + GPT-4o (D) | 0.799 | 0.779 | 0.863 | 0.860 |
| GPT-4o (G) + GPT-3.5 (D) | 0.879 | 0.878 | 0.924 | 0.926 |
| Qwen2.5 (G) + GPT-4o (D) | 0.772 | 0.742 | 0.849 | 0.832 |
| GPT-4o (G) + Qwen2.5 (D) | 0.879 | 0.871 | 0.916 | 0.919 |
| Llama 3 (G) + GPT-4o (D) | 0.846 | 0.835 | 0.879 | 0.875 |
| mLLMNego (GPT-4o (G) + Llama 3 (D)) | **0.886** | **0.881** | **0.925** | **0.927** |

（3）在mLLMNego架构中，GPT-4o作为生成器与Llama 3作为鉴别器的组合，在两个数据集上的表现都达到了最佳水平，显著高于其他模型组合。例如，在Weibo16数据集上，mLLMNego的$F1$值比其他架构高出0.1%到14.1%；在Twitter15数据集上，mLLMNego的$F1$值比其他架构高出0.3%到15%。通过合理选择生成器和鉴别器，不仅能够

优化生成器的输出，还能通过鉴别器的辅助提升整体推理的精确度和鲁棒性。

### 5.6　消融实验

在验证了自适应上下文构造策略和多大语言模型协商机制的有效性的基础上，本节消融分析重点探讨了所提CollabDetection方法中“多维度证据”和“特征融合模块”的贡献。如图11所示，设计了六个变种：(1)w/o evidence：仅使用文本特征提取模块来学习帖子的表示，不使用mLLMNego生成的多维度证据特征。(2)Concat：在融合模块中，简单地将文本特征和多维度证据特征进行拼接。(3)加权和(Weighted_sum)：根据文本特征和多维度证据特征的相对重要性，为两者分配不同的权重后进行加权和融合。(4)SVM：采用支持向量机模型来融合文本表示和多维度证据特征。(5)LR：使用逻辑回归来结合文本特征和多维度证据特征。(6)Attention：采用注意力机制融合特征。后五个变种方法的设计旨在与本文提出的MFB融合策略进行比较，以验证MFB的有效性。

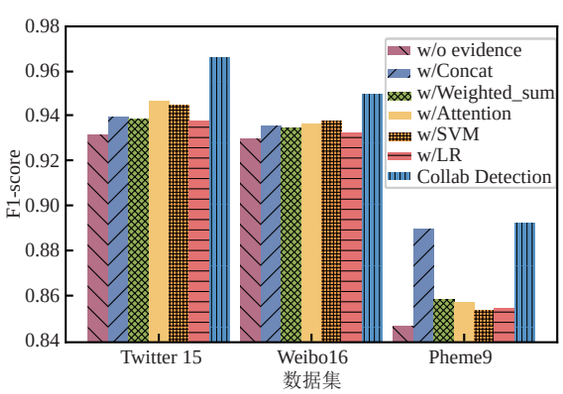


图11　CollabDetection消融实验性能对比

从图11可以观察到，与不使用证据特征的w/o evidence方法相比，所有引入证据特征的变种方法在三个数据集上的$F1$分数均有所提升，表明由mLLMNego方法生成的多维度证据特征在虚假新闻检测中具有重要作用，去除多维度证据特征会导致性能大幅下降。在所有引入证据特征的变种方法中，采用MFB融合策略的CollabDetection方法在三个数据集上均取得了最高的$F1$分数，显著优于所有对照方法，验证了CollabDetection模型设计的合理性和有效性。

### 5.7　案例分析

本文通过两个典型案例，分析基于多大语言模型协商的多维度证据生成方法mLLMNego在虚假新闻检测中的表现。图12和图13分别展示了测试集中的两个典型案例，由于篇幅限制对部分回答内容进行了省略。

案例1：mLLMNego预测正确，GPT3.5零样本预测错误
帖子内容：
[2013上海国际足球邀请赛.回放#上海申花0:1大田市民#]2013上海国际足球邀请赛开战，上海申花坐镇虹口足球场0-1不敌大田市民。申花全替补出战，投资人@朱骏身披18号球衣首发出战并打满全场，下半场朱骏曾补射打入一球，但被判越位无效。大田市民在第50分钟由韩德熙打入制胜一球。http://t.cn/zHpnFHk
mLLMNego回答：
Real news. The tweet provides a factual account of a sports event, specifically the 2013 Shanghai International Football Invitational Tournament. It includes specific details such as the teams involved (Shanghai Shenhua and Daejeon Citizen), the venue (Hongkou Football Stadium), and the outcome (0-1 loss for Shanghai Shenhua). The mention of Zhu Jun, an investor, playing in the match and having a goal disallowed for offside adds to the completeness of the information. The source appears reliable, as it provides a URL for further verification. The emotional tone is neutral, focusing on reporting the match's events. Hashtags are used to categorize the content, and the intent is to inform about the match's outcome. The topic is sports, making it unlikely to be fake news.
GPT3.5零样本回答：
Fake news. The event described, the 2013 Shanghai International Football Invitational, is outdated and not recent. The lack of recent context or relevance raises doubts about the accuracy of the information. Additionally, the emotional tone in the tweet is neutral and lacks urgency or credibility... The use of hashtags and the intent behind the tweet also suggest a possible nostalgic or fictional narrative rather than a current event.
真实性标签：
真实新闻

图12　多维度证据生成案例分析

图12展示了mLLMNego与GPT-3.5在同一微博中的不同表现。该推文涉及2013年上海国际足球邀请赛的比赛结果，mLLMNego正确判定为“真实新闻”，GPT-3.5错误判定为“虚假新闻”。两者在基本信息和情感识别上相似，但mLLMNego在细节完整性、来源可靠性和证据推理等方面明显优于GPT-3.5。mLLMNego提供了详细信息，如参赛球队、比赛场地、比赛结果等，增强了信息的可信度，并指出了可验证的来源链接和相关话题标签。相比之下，GPT-3.5对推文的具体细节理解不足，容易依赖表面信息，导致误判。该案例表明，mLLMNego在信息整合和证据生成方面具有显著优势。

图13中的案例则进一步展示了mLLMNego方法在多轮协商中的动态调整和自我修正能力。该案例分析了一条关于高中生午休时间炒股获利7200万美元的推文，初始判断存在偏差，经过多轮协商，mLLMNego将其判定为“虚假”。在初轮判断中，模型基于表面信息错误判定为“真实新闻”。经过第二轮质疑，模型审视推文的合理性，指出如此巨

额的短期盈利缺乏可信证据支持，且RTLNy不是广泛认可的新闻来源，标签本身缺乏权威性。经过三轮协商，模型综合多维度证据，修正判断为“虚假新闻”。这一过程展现了mLLMNego的动态优化能力，能够有效弥补初始证据的不足。

案例2：mLLMNego多轮协商后预测正确
帖子内容：#rtlny: because a stuyvesant senior made $72 million trading stocks on his lunch break URL URL
第一轮回答：
Real news. The tweet provides specific details about the student's achievement, including the amount of money made ($72 million) and the school he attends (Stuyvesant High School)...The tweet is attributed to a reputable news source, RTLNy, which increases the credibility of the information... The hashtag #rtlny suggests the information might come from a legitimate news channel, enhancing the perception of accuracy. The intent of the tweet appears to be informative, rather than to deceive or manipulate...
第二轮回答：
No. The claim about a Stuyvesant High School senior making $72 million trading stocks on his lunch break is implausible and unverifiable. There is no reliable evidence supporting such an extraordinary claim. The source credibility is questionable, as RTLNy is not a widely recognized news outlet, and the hashtag #rtlny lacks association with any verifiable media organization... The tweet's sensationalized nature and focus on attracting attention make it more likely to be fake news rather than a factual statement...
第三轮回答：
Yes. The discriminator's analysis highlights several red flags that suggest the tweet is likely fake news. Firstly, the claim of a Stuyvesant senior making $72 million in a short period is an unlikely scenario and lacks supporting proof... Additionally, the source (RTLNy) lacks credibility, as it is not a recognized news platform, and the associated hashtag #rtlny offers no verifiable trace to a legitimate source. The exaggerated nature of the statement suggests it aims to mislead or sensationalize, which is a common characteristic of fake news.
真实性标签：
虚假新闻

图13 多轮协商案例分析

## 6 总 结

虚假新闻检测作为一种自动化检测手段，能够为人工审核提供有效的辅助手段，通过初步筛查和标记潜在虚假新闻，从而提高整体检测效率并减轻人工审核负担。本文聚焦于虚假新闻检测领域中传统证据增强方法存在的检索噪声问题，提出了基于证据生成和语义融合的两阶段虚假新闻检测方法CollabDetection。在证据生成阶段，设计了自适应上下文构造策略和多大语言模型协商机制，通过引入生成器和鉴别器进行多轮对抗协商，模拟人类协作决策过程，确保证据文本的准确性。在语义融合阶段，以小语言模型高效的特征表示能力为基础，结合大语言模型生成的多维度证据，采用多模态因子分解双线性池化技术聚合多语义特征。本文选取了19个模型进行对比实验，分别对自适应上下文构造策略和协商机制的有效性进行了深入分析，并做了充分的消融实验。基于真实数据集的实验结果表明，所提CollabDetection方法在准确性和泛化能力上均具有显著优势，能够稳定地生成高质量的多维度证据，从而提升虚假新闻检测的效果。

然而，本文仍然存在一些不足：(1)对于基于多大语言模型协商的多维度证据生成方法，本文目前仅考虑了七个维度的证据，这可能限制了模型对复杂语境的全面覆盖。未来研究可以尝试引入更多维度的证据，例如传播路径以及用户信息等。(2)对于融合多特征语义的虚假新闻检测方法，本文采用了多模态因子分解双线性池化技术进行特征融合，但尚未深入研究其他潜在的融合策略。(3)虚假新闻通常具有动态演化和传播迅速的特点，而本文的方法主要基于静态数据集进行实验，尚未结合实时数据流进行动态更新。

## 参 考 文 献


[1] Dou Y, Shu K, Xia C, et al. User preference-aware fake news detection//Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. Virtual , Canada, 2021: 2051-2055

[2] Min E, Rong Y, Bian Y, et al. Divide-and-Conquer: Post-User Interaction Network for Fake News Detection on Social Media//Proceedings of the ACM Web Conference 2022. Lyon, France, 2022: 1148-1158

[3] Ma J, Gao W, Mitra P, et al. Detecting rumors from microblogs with recurrent neural networks//Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence. New York, USA, 2016: 3818-3824

[4] Liu Nan, Zhang Feng-Li, Yin Jia-Qi, et al. Edge inference-enhanced contrastive learning for social media rumor detection. Computer Science, 2023, 50(11): 49-54 (in Chinese)
(刘楠, 张凤荔, 尹嘉奇, 等. 基于边推断增强对比学习的社交媒体谣言检测模型. 计算机科学, 2023, 50(11): 49-54)

[5] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA, 2017: 6000-6010

[6] Devlin J, Chang M-W, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics. Minneapolis, USA, 2019: 4171-4186

[7] Achiam J, Adler S, Agarwal S, et al. GPT-4 Technical Report. arXiv preprint arXiv:2303.08774, 2023

[8] Asher N, Bhar S, Chaturvedi A, et al. Limits for learning with language models//Proceedings of the 12th Joint Conference on Lexical and Computational Semantics. Toronto, Canada, 2023: 236-248

[9] Zhong Jiang, Gao Jin-Peng, Huang Jing-Wang, et al. Evidence-enhanced and local semantic interaction-based multimodal fake news detection. Chinese Journal of Computers, 2024, 48(03): 556-571 (in Chinese)
(钟将, 高晋鹏, 黄敬旺等. 基于证据增强和局部语义交互的多模态虚假新闻检测. 计算机学报, 2024, 48(03): 556-571)

[10] Hu X, Guo Z, Chen J, et al. MR2: A benchmark for multimodal retrieval-augmented rumor detection in social media//Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. Taipei, China, 2023: 2901-2912

[11] Castillo C, Mendoza M, Poblete B. Information credibility on Twitter//Proceedings of the 20th International Conference on World Wide Web, Hyderabad, India, 2011: 675-684

[12] Nan Q, Cao J, Zhu Y, et al. MDFEND: Multi-domain Fake News Detection//Proceedings of the 30th ACM International Conference on Information & Knowledge Management. Virtual Australia, 2021: 3343 - 3347

[13] Huang Xue-Jian, Wang Gen-Sheng, Luo Yuan-Sheng, et al. Weibo rumors real-time detection model based on fusion of multi user features and content features, Journal of Chinese Computer Systems 2022, 43(12): 2518-2527 (in Chinese)
(黄学坚, 王根生, 罗远胜, 等. 融合多元用户特征和内容特征的微博谣言实时检测模型. 小型微型计算机系统, 2022, 43(12): 2518-2527)

[14] Yue Z, Zeng H, Zhang Y, et al. MetaAdapt: Domain adaptive few-shot misinformation detection via meta learning//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. Toronto, Canada, 2023: 5223-5239

[15] Ma J, Dai J, Liu Y, et al. Contrastive learning for rumor detection via fitting beta mixture model//Proceedings of the 32nd ACM International Conference on Information and Knowledge Management. Birmingham, UK, 2023: 4160-4164

[16] Nan Q, Sheng Q, Cao J, et al. Let silence speak: Enhancing fake news detection with generated comments from large language models//Proceedings of the 33rd ACM International Conference on Information and Knowledge Management. Boise, USA, 2024: 1732-1742

[17] Ke Jing, Xie Zhe-Yong, Xu Tong, et al. Implicit semantic enhancement-based fine-grained fake news detection using large language models. Journal of Computer Research and Development, 2024, 61(05): 1250-1260 (in Chinese)
(柯婧, 谢哲勇, 徐童, 等. 基于大语言模型隐含语义增强的细粒度虚假新闻检测方法. 计算机研究与发展, 2024, 61(05): 1250-1260)

[18] Lu Y, Bartolo M, Moore A, et al. Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Dublin, Ireland, 2022: 8086-8098

[19] Wang B, Ma J, Lin H, et al. Explainable Fake News Detection with Large Language Model via Defense Among Competing Wisdom//Proceedings of the ACM Web Conference 2024. Singapore, 2024: 2452-2463

[20] Hu B, Sheng Q, Cao J, et al. Bad Actor, Good Advisor: Exploring the Role of Large Language Models in Fake News Detection//Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada, 2024: 22105-22113

[21] Popat K, Mukherjee S, Yates A, et al. DeClarE: Debunking Fake News and False Claims using Evidence-Aware Deep Learning//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium, 2018: 22-32

[22] Liao H, Peng J, Huang Z, et al. MUSER: A multi-step evidence retrieval enhancement framework for fake news detection//Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Long Beach, USA, 2023: 4461-4472

[23] Xu W, Wu J, Liu Q, et al. Evidence-aware fake news detection with graph neural networks//Proceedings of the ACM Web Conference 2022. Lyon, France, 2022: 2501-2510

[24] Zhang X, Cao J, Li X, et al. Mining dual emotion for fake news detection//Proceedings of the Web Conference 2021. Ljubljana, Slovenia, 2021: 3465-3476

[25] Zhou X, Shu K, Phoha V V, et al. "This is fake! shared it by mistake": Assessing the intent of fake news spreaders//Proceedings of the ACM Web Conference 2022. Lyon, France, 2022: 3685-3694

[26] Hu L, Yang T, Zhang L, et al. Compare to The Knowledge: Graph Neural Fake News Detection with External Knowledge//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Bangkok, Thailand, 2021: 754-763

[27] Grünwald P D. The minimum description length principle. Cambridge USA: MIT press, 2007

[28] Yu Z, Yu J, Fan J, et al. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 2017: 1821-1830

[29] Ma J, Gao W, Wong K-F. Detect Rumors in Microblog Posts Using Propagation Structure via Kernel Learning//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, Canada, 2017: 708-717

[30] Kochkina E, Liakata M, Zubiaga A. All-in-one: Multi-task Learning for Rumour Verification// Proceedings of the 27th International Conference on Computational Linguistics. Santa Fe, USA, 2018: 3402-3413

[31] Kim Y. Convolutional Neural Networks for Sentence Classification//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar, 2014: 1746-1751

[32] Liu Y, Ott M, Goyal N, et al. RoBERTa: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692, 2019

[33] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners// Advances in Neural Information Processing Systems. Virtual, 2020: 1877-1901

[34] Touvron H, Martin L, Stone K, et al. Llama 2: Open

foundation and fine-tuned chat models. arXiv preprint arXiv: 2307.09288, 2023

[35] Dubey A, Jauhri A, Pandey A, et al. The llama 3 herd of models. arXiv preprint arXiv:2407.21783, 2024

[36] He Z, Li C, Zhou F, et al. Rumor detection on social media with event augmentations// Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. Virtual , Canada, 2021: 2020-2024

[37] Zhang W, Zhong T, Li C, et al. CausalRD: A causal view of rumor detection via eliminating popularity and conformity biases// IEEE International Conference on Computer Communications. London, UK, 2022: 1369-1378

[38] Ma J, Liu Y, Liu M, et al. Curriculum contrastive learning for fake news detection// Proceedings of the 31st ACM International Conference on Information & Knowledge Management. Atlanta, USA, 2022: 4309-4313

[39] Huang Z, Lv Z, Han X, et al. Social bot-aware graph neural network for early rumor detection// Proceedings of the 29th International Conference on Computational Linguistics. Gyeongju, Republic of Korea, 2022: 6680-6690

[40] Sun T, Qian Z, Dong S, et al. Rumor detection on social media with graph adversarial contrastive learning// Proceedings of the ACM Web Conference 2022. Lyon, France, 2022: 2789-2797

[41] Yang C, Zhang P, Gao H, et al. Deciphering Rumors: A multi-task learning approach with intent-aware hierarchical contrastive learning//Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Miami, USA, 2024: 4471-4483

**MA Xiao**, Ph. D., associate professor. Her research interests include recommender systems, fake news detection, graph representation learning, data mining and so on.

**LI Xin-Yu**, M. S. candidate. Her research interest is fake news detection.

**ZENG Jiang-Feng**, Ph. D., associate professor. His research interests include fake news detection, affective computing, recommender systems, artificial intelligence and so on.

**Background**

In the information age, the proliferation of fake news, prioritizing emotion and imagination over facts and truth, poses significant risks to media credibility and social stability. Detecting fake news remains a critical and ongoing research challenge.

Existing machine learning-based methods can be predominantly grouped into content-based and context-based. Context-based methods trace the origin of fake news by analyzing the pathways of information dissemination and the network structure. However, these methods often suffer from information delay, making it difficult to meet the demand for early detection of fake news. Content-based methods, on the other hand, focus on the semantic understanding of the news text itself, aiming to identify fake news as early as possible by directly analyzing content features. With the development of semantic representation techniques based on Transformer, the performance of content-based methods has been significantly enhanced. However, relying exclusively on the textual content of news itself to determine its authenticity is insufficient since the authenticity of news is usually supported by abundant evidence and precise facts.

To this end, this paper proposes an innovative two-stage approach called CollabDetection that combines evidence generation and semantic fusion. In the stage of evidence generation, it utilizes large language models (LLMs) to generate potential evidence, mitigating the noise introduced by traditional evidence retrieval methods. In the stage of semantic fusion, evidence and news content are first encoded by a pre-trained small language model and then aggregated by utilizing the multi-modal factorized bilinear pooling technique. Experiments on the Twitter15, Weibo16 and Pheme9 datasets demonstrate that CollabDetection significantly outperforms baseline models by leveraging the strengths of both large and small language models in a collaborative framework.

The broader significance of this project lies in its potential to mitigate the spread of fake news on platforms such as Weibo and Twitter, thereby fostering a healthier and more transparent digital information ecosystem.