徽标, 公司名称

描述已自动生成

**2025年（第18届）**

**中国大学生计算机设计大赛**

人工智能实践赛作品报告

作品编号：

作品名称：

填写日期：

**目 录**

[第1章 作品概述 1](#_Toc195631183)

[1.1 研究背景 1](#_Toc195631184)

[1.2 主要功能 1](#_Toc195631185)

[1.2.1 虚假文本检测 1](#_Toc195631186)

[1.2.2 虚假图文检测 2](#_Toc195631187)

[1.2.3 虚假视频检测 2](#_Toc195631188)

[1.3 创新点 2](#_Toc195631189)

[1.3.1 SDSA 数据去冗余与语义对齐机制 2](#_Toc195631190)

[1.3.2 BMLHF 多模态平衡融合机制（MIB） 2](#_Toc195631191)

[1.3.3 FEDPM 联邦多模态双通道学习框架 3](#_Toc195631192)

[1.4 应用价值与推广前景 3](#_Toc195631193)

[第2章 问题分析 3](#_Toc195631194)

[2.1 问题来源 3](#_Toc195631195)

[2.2 现有解决方案 4](#_Toc195631196)

[2.2.1 数据去冗余与语义对齐 4](#_Toc195631197)

[2.2.2 多模态平衡融合 4](#_Toc195631198)

[2.2.3 联邦式多模态协同建模 5](#_Toc195631199)

[2.3 本作品要解决的痛点问题 5](#_Toc195631200)

[2.3.1 多模态数据冗余与语义不一致 5](#_Toc195631201)

[2.3.2 融合过程模态失衡 5](#_Toc195631202)

[2.3.3 分布式数据难以统一建模 6](#_Toc195631203)

[2.4 解决问题的思路 6](#_Toc195631204)

[第3章 技术方案 7](#_Toc195631205)

[3.1 系统架构 7](#_Toc195631206)

[3.2 基于语义蒸馏和结构相似性的三重对齐网络 7](#_Toc195631207)

[3.3 基于动态权重分配的多模态信息平衡技术 7](#_Toc195631208)

[3.4 基于个性化调制器的LoRA大模型微调方法 9](#_Toc195631209)

[第4章 系统实现 9](#_Toc195631210)

[4.1 数据来源与预处理流程 10](#_Toc195631211)

[4.2 多模态信息平衡融合机制的实现 11](#_Toc195631212)

[第5章 测试分析 12](#_Toc195631213)

[第6章 作品总结 14](#_Toc195631214)

[6.1 作品特色与创新点 14](#_Toc195631215)

[6.2 应用推广 14](#_Toc195631216)

[6.3 作品展望 15](#_Toc195631217)

[参考文献 16](#_Toc195631218)

# 作品概述

## 研究背景

随着通信技术与数字技术的迅猛发展与广泛应用，风险内容在网络空间中呈现出快速增长的趋势，成为当前社会管理与网络空间治理面临的重要挑战。多模态风险内容频繁出现于诈骗短信与电话、虚假新闻传播、伪造人脸图像与音视频等形式中。这些风险内容通常来源于多种模态媒介，包括文本、图像、音频、视频及流数据等，不仅形式多样、体量庞大，而且真假难辨，给识别与治理带来了显著难度。

在实际应用中，风险内容往往以混合模态出现，增加了传统识别技术的复杂性。因此，如何构建高质量的多模态风险内容数据集、有效应用并微调大模型、同时提取各模态数据中的关键特征并保留其语义信息，成为该领域亟待攻克的核心难题。为应对这一挑战，有必要围绕诈骗识别、伪造视频与音频检测、虚假新闻识别等典型多模态风险识别任务，开展系统性的研究与技术应用，为网络空间治理提供有力支撑。

## 主要功能

本项目围绕多模态风险内容识别任务，设计并实现了以下三个核心功能，充分融合了多项关键技术创新，旨在提升虚假内容检测的准确性、鲁棒性与实用性：

### 虚假文本检测

通过构建高质量的文本风险数据集，结合SDSA数据去冗余与对齐机制，对原始文本数据进行清洗、提纯与语义对齐，有效剔除冗余与噪声信息。借助BMLHF（MIB）多模态特征融合机制，在保留语义表达完整性的同时，实现对文本中的潜在虚假信息（如诈骗话术、捏造信息、情绪操控等）的高精度识别，具备良好的泛化能力。

### 虚假图文检测

面对图文混合场景中虚假内容的复杂特性，系统采用SDSA模块对图文内容分别进行结构化清洗与语义对齐，建立图文间跨模态关联。通过BMLHF（MIB）多模态融合机制，对图像特征与文本信息进行深度语义融合，增强模型对图文关系合理性、内容一致性的判断能力。结合FEDPM模型进行图文协同学习与判别，有效识别伪造图片配合虚假文本进行信息操控的风险行为。

### 虚假视频检测

针对伪造视频（如Deepfake、剪辑合成）带来的信息安全风险，项目引入FEDPM（联邦多模态表示学习）模型，实现对视频模态中的图像帧、语音轨道、字幕文本等多源信息的联合建模与平衡融合。通过BMLHF（MIB）机制处理各模态间特征异构性，提升模型对合成内容、语音仿冒、人脸伪造等复杂欺诈行为的识别精度。同时，结合SDSA数据清洗机制，有效筛除低质量样本，提升训练数据的有效性与模型鲁棒性。

## 创新点

### SDSA 数据去冗余与语义对齐机制

针对当前多模态数据在采集过程中普遍存在数据冗余严重、噪声比例高、语义不一致等问题，项目提出SDSA（Selective De-redundancy and Semantic Alignment）数据去冗余与语义对齐机制，该机制为虚假内容识别提供了更高质量、更一致性的输入数据，有效提高了模型训练效率与识别准确率。

### BMLHF 多模态平衡融合机制（MIB）

针对多模态信息融合过程中“强模态主导、弱模态退化”等常见问题，项目设计并实现了BMLHF（Balanced Multi-modal Learning with Hybrid Fusion）多模态平衡融合机制，该机制显著提升了模型对多模态输入的综合理解与决策能力，特别适用于复杂场景下的虚假图文与视频识别任务。

### FEDPM 联邦多模态双通道学习框架

为提升模型在数据分布广泛、隐私保护要求高的实际环境下的可部署性与泛化能力，项目提出FEDPM（Federated Dual-path Multi-modal Learning）联邦多模态双通道学习框架，该框架突破了传统多模态建模对集中数据依赖的限制，在保障隐私的同时，显著提升了模型的跨域泛化能力与工程可落地性。

## 应用价值与推广前景

本项目聚焦虚假文本、图文及视频内容识别，构建了面向多模态风险内容的智能识别体系，具有显著的实际应用价值和广阔的推广前景。项目成果可广泛应用于网络安全治理、内容审核、媒体舆情监测等关键领域，助力政府监管部门提高对虚假信息的识别效率与干预能力，强化社会治理数字化水平。同时，可为社交平台、新闻客户端、视频平台等内容发布方提供高精度内容安全工具，提升内容合规水平与用户信任度。

项目采用模块化架构，支持多系统集成，具备良好的可部署性与扩展性；联邦式建模机制保障数据安全合规，适用于多组织、跨行业的协同应用场景。随着AIGC等生成内容的快速发展和虚假信息传播手段不断演化，社会各界对多模态识别技术的需求日益增长，项目具备良好的技术引领性和商业转化潜力，推广前景广阔。

# 问题分析

## 问题来源

随着大模型技术的快速发展和网络数据的爆炸式增长，文本、图像、音频、视频等多模态信息在互联网中高速传播。与此同时，虚假文本、伪造图像、合成语音与深度伪造视频等风险内容不断涌现，形成更具欺骗性和传播力的多模态虚假信息，严重威胁网络安全与社会信任。

当前主流的风险识别技术仍以单模态或浅层融合为主，难以处理多模态数据间的语义对齐、模态失衡和跨源异构等复杂问题，且普遍缺乏对高质量数据筛选与隐私合规的系统支持，识别效果受限、泛化能力不足。

因此，亟需构建面向多模态风险内容的大模型识别技术体系，从数据预处理、特征融合到协同建模形成全流程能力，突破当前识别能力瓶颈，满足复杂风险场景下的智能识别需求。

## 现有解决方案

近年来，多模态风险内容检测逐渐成为人工智能与网络安全领域的研究热点，相关研究主要集中在数据清洗与去冗余、模态融合策略、多源建模机制等方向。然而，现有技术仍存在诸多局限，难以满足实际复杂场景中对高质量数据处理、模态平衡理解及隐私合规建模的综合需求。

### 数据去冗余与语义对齐

现有多模态数据清洗方法多采用通用的数据增强、重复样本识别（如LSH hashing、聚类去重）或噪声过滤策略（如置信度筛选、简单规则清洗），如 CLIP-Lite[1]、BLIP[2]等模型在预训练阶段引入图文匹配得分来辅助数据筛选。但这些方法多为静态规则或单模态驱动，难以应对多模态数据中潜藏的语义偏移与伪相关性。

在语义对齐方面，常见做法包括使用对比学习（如ALBEF[3]、VLMo）或多任务目标实现模态嵌入对齐，但对于“弱语义”模态（如截图图像、合成语音）表现不稳定。此外，大部分方法仅对齐图文模态，缺乏统一处理图、文、音、视频的能力，无法满足复杂风险内容对高一致性语义建模的需求。

### 多模态平衡融合

当前多模态融合方法以“Early Fusion”（特征级融合）和“Late Fusion”（决策级融合）为主，部分方法如MMBT、UNITER尝试引入统一表示空间，提升融合深度。但在多模态内容质量不平衡或模态缺失情况下，常出现“强模态主导、弱模态失效”的问题，导致模型泛化能力下降。

为应对模态不平衡问题，一些研究引入门控机制（如Modality Dropout）或模态注意力（如ViLT、METER）进行模态调控，但仍缺乏对动态内容质量变化的自适应建模能力。此外，现有方法多数关注特定任务（如图文匹配、VQA），泛用性与可扩展性不足，难以适配诈骗检测、伪造识别等复杂场景下的多任务需求。

### 联邦式多模态协同建模

针对数据隐私与跨域建模问题，已有部分研究将联邦学习（FL）引入多模态任务场景，如FedAvg、FedMA等方法被用于跨机构图文模型协同训练。但传统FL方法多聚焦于结构一致、模态单一的数据源，面对异构多模态、跨模态不对齐等问题仍缺乏有效支持。

近年来，少量研究如 FedGL[4]尝试在图神经网络中引入模态间建模机制，部分研究（如FedVQA[5]）对多模态VQA任务进行尝试性迁移，但仍面临以下问题：一是无法同时建模模态内和模态间的复杂关系，二是对数据质量与模态缺失情况处理能力较弱，三是缺乏动态适应不同客户端模态配置的能力。

## 本作品要解决的痛点问题

### 多模态数据冗余与语义不一致

在多模态虚假内容识别任务中，数据往往来源复杂、模态多样，尤其在网络环境下采集的图像、文本、音频、视频等信息中，存在大量冗余片段、模糊标签和语义不一致的问题。这不仅降低了训练数据的有效性，还可能引入误导性特征，影响模型的学习效率和识别准确率。现有方法多依赖单模态静态清洗策略，难以对多模态语义关系进行统一建模与优化。项目提出 SDSA 数据去冗余与语义对齐机制，从多模态语义一致性的角度出发，对输入数据进行筛选、提纯与对齐，为模型提供更高质量的输入支撑，有效缓解该问题。

### 融合过程模态失衡

当前多模态融合方法普遍存在模态主导性强弱不均的现象。在图文、音视频等融合场景中，主导模态（如图像或视频）在训练过程中会压制弱模态（如文本或语音）的表达能力，导致弱模态特征被忽视，信息融合不充分，最终影响模型的鲁棒性和泛化能力。为应对这一问题，项目设计并实现了 BMLHF 多模态平衡融合机制（MIB），引入模态重要性动态调节策略和多路径融合结构，实现各模态之间的协同学习和语义互补，从而提升模型对复杂场景下多模态虚假内容的识别效果。

### 分布式数据难以统一建模

在多源、多场景的实际应用环境中，数据分布存在显著差异，且受限于隐私保护法规与数据孤岛问题，传统集中式多模态模型训练方法难以满足真实部署需求。此外，风险内容的生成机制、模态组合方式和攻击手法具有高度异质性，导致模型迁移性差、适应能力不足。针对上述问题，项目提出 FEDPM 联邦多模态双通道学习框架，结合联邦学习与多模态建模策略，在保护数据隐私的同时实现分布式模型协同优化，显著提升模型的跨场景泛化能力与工程可部署性。

## 解决问题的思路

本项目旨在通过构建融合SDSA、BMLHF与FEDPM机制的多模态风险内容识别系统，提升模型在多源异构数据环境下的识别精度、处理效率及跨场景适应能力。系统面向**文本、图像与视频**共三种主要模态，实现虚假文本检测、图文一致性识别与伪造视频识别等核心功能。

在功能与性能层面，本系统在真实环境模拟测试中表现出良好的计算效率和识别能力。基于100次运行测试，模型平均响应时间为175ms，整体内存开销控制在边缘部署可接受范围内，适合联邦学习等分布式场景下的实际应用。在风险内容识别方面，系统在多个标准测试集上实现了96%的准确率，F1分数达到0.83，充分体现了其在精度与鲁棒性方面的优势。

在实验数据方面，项目选用三个常用的多模态虚假信息检测公开数据集：Twitter、Weibo 和 Fakeddit，覆盖中英文社交平台、多领域事件，具有良好的代表性与挑战性。Twitter 数据集包含992条真实新闻和1101条假新闻，样本采集时间跨度从2013年至2017年，涵盖社会、灾难、政治等主题，每条样本均包含文本、图像与真假标签。Weibo 数据集是中文多模态任务中最具代表性的语料之一，由4700条真实新闻和4000条假新闻组成，来源于新浪微博，领域覆盖广泛。Fakeddit 数据集则来自Reddit平台，我们选取其中一个包含文本与图像的二分类子集进行实验，该子集包含约60,000条带有真假标签的样本，具有更大规模与复杂语义的特点。

# 技术方案

【填写说明：从原理层面，详细介绍系统所采用的技术方案，先总体介绍，给出技术路线框架图，然后分模块详细介绍。着重介绍解决问题的思路，以及所涉及的模型、协议、算法等，以及可能的对算法的改进；原创工作详述，非原创工作简述，并尽可能标注引用文献】

## 系统架构

## 基于语义蒸馏和结构相似性的三重对齐网络

### 总体介绍

多模态信息冗余和结构信息丢失问题本质上是跨模态交互中因噪声干扰和特征对齐失准引发的有效信息筛选与语义结构完整性失衡。为解决这一问题，我们提出基于语义蒸馏和结构相似性的三重对齐网络（以下简称SDSA），通过分层语义蒸馏和跨模态相似性度量，动态对齐多模态表征，自适应过滤冗余信息以保留结构特征，系统分为两大部分：

1. **语义蒸馏**：通过蒸馏模态特定的特征，保留与假新闻检测任务相关的语义信息，同时消除冗余信息。

2.**结构对齐**：确保在模态内、模态间以及跨模态保留结构信息（邻域关系）。

整体架构的框架图如图3.1所示：

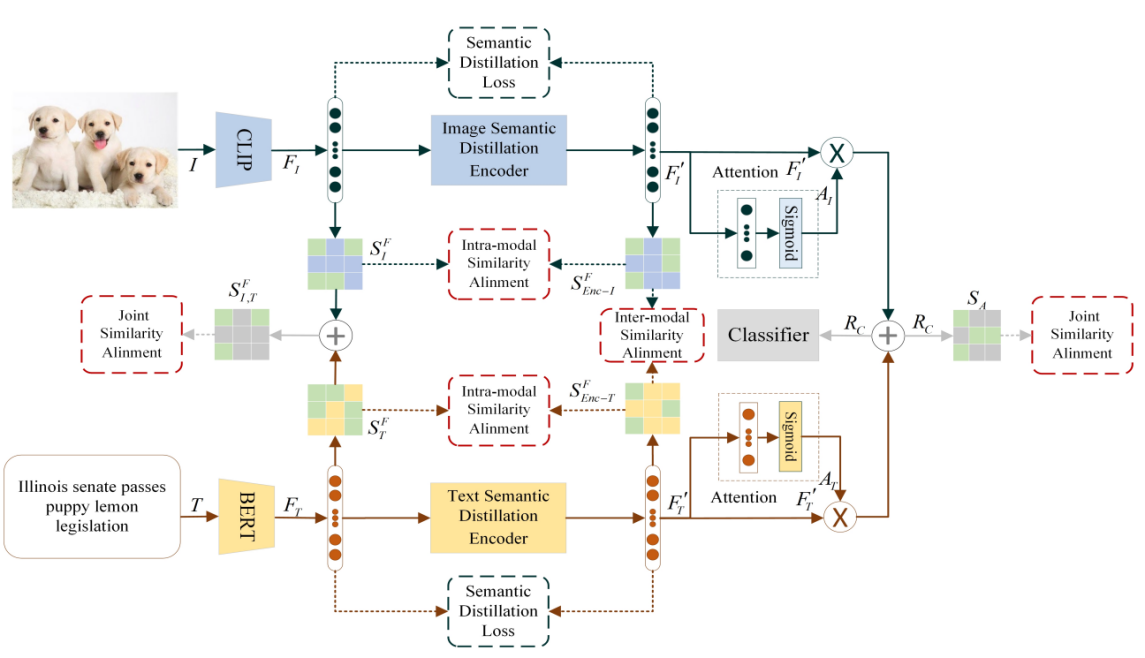


图3.1 SDSA方法总体架构图

### 技术方案：系统组件

#### 语义蒸馏

语义蒸馏模块基于信息瓶颈理论对输入的图像和文本特征进行编码，从而保留与假新闻检测任务相关的语义信息，去除不相关的冗余信息。数学上表示为：

其中代表与任务相关的语义信息，代表冗余信息。

目标是最小化（与任务相关的语义信息）与之间的差异，这通过最小化KL散度来实现。

#### 三重相似度对齐

为了保留输入数据的结构信息，我们提出了一个三重相似度对齐模块，其分为三个部分：

模态内相似度对齐：对每个模态（图像和文本）内的相似度矩阵进行对齐，确保编码前后的邻域结构一致。

模态间相似度对齐：对两个模态（文本和图像）之间的相似度矩阵进行对齐，最小化它们之间的差异，从而减少模态间的不同。

联合相似度对齐：对融合后的特征进行对齐，保留不同模态融合特征的结构完整性。通过加权方法平衡不同模态结构信息的贡献。

整体相似度损失函数为：

其中是分类损失，是结构相似度损失，是信息瓶颈损失。

#### 分类

最终步骤是通过交叉熵损失对融合后的特征进行分类，将样本标记为假新闻或真实新闻。这是在特征融合和相似度对齐步骤之后进行的。

## 基于动态权重分配的多模态信息平衡技术

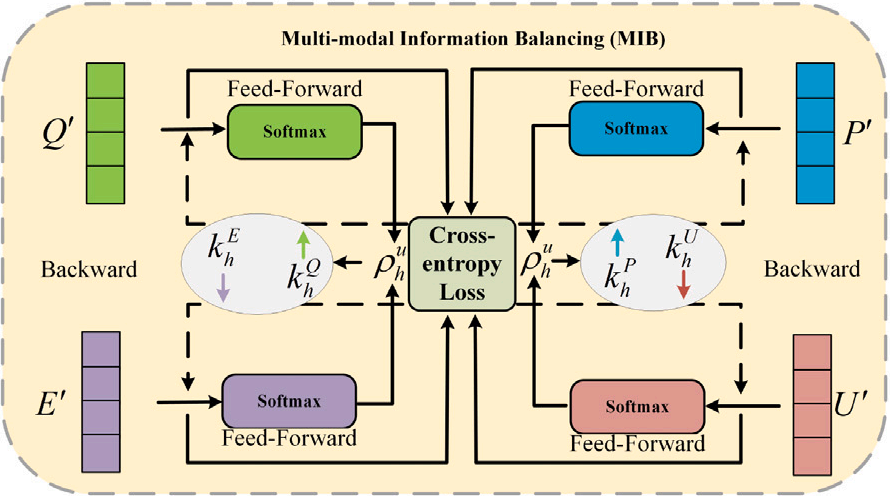


图3.1 多模态信息平衡模块架构图

模态不平衡问题是指在优化过程中，主导模态会抑制其他模态性能的现象。为解决这一问题，我们设计了多模态信息平衡（MIB）模块，在各模态的优化过程中动态地为它们分配权重，自适应地调节模型的优化过程，以实现模态平衡，如图1所示。

对于的每个通道，其参数为，梯度更新过程由下式表示：

其中是第次迭代更新后的参数。

我们通过更精细地监测信息，自适应地调整每个模态的梯度。这些信息可以表述为每个模态的特定信息，对数几率得分（logit score）能直接反映不同模态的激活程度。因此，在本文中，我们计算每个模态的模式视图和语义视图的对数几率得分，以探究每个模态的信息量。对数几率得分包含，其中和是的参数，表示中的第个特征，是第次迭代中的一个随机批次。我们定义来进一步量化不同模态视图的信息：

其中 ，是类别数。

我们设计差异率来衡量各模态对优化过程的影响：

越大，与其他模态视图相比，相应模态视图中的信息量就越大，从而导致模态不平衡。

信息量较大的模态在模型的优化过程中起着主导作用[7][8][9]，并抑制其他模态的优化过程。因此，我们设计平衡因子来平衡各模态的优化过程。

然后，我们将整合到公式 (1) 中。的更新过程如下：

通过使用，我们抑制了性能较好的模态视图的优化，而性能较差的模态视图则不受影响。

在获取多模态的融合特征后，我们将其输入到一个分类器中，该分类器包含一层多层感知器（MLP）和 激活函数，以获得预测标签：

其中是分类器，是分类器的参数。

为了增强模型的分类能力，我们最小化交叉熵损失：

为了增强特征的判别能力，我们采用跨模态监督对比损失。是中与属于同一类别的索引集，是中与类别不同的索引集。

其中，是的基数，是一个标量参数，根据经验设置为。因此，总损失可以表示为：

其中，是一个平衡因子。

我们使用 Adam 优化器在总损失下寻找最优的网络参数。对于测试样本集 ，我们使用最优参数、、、、、来获取相应的特征，然后将这些特征输入到分类器中以获得标签。通过我们的方法，可以调节每个模态的优化过程，有效缓解模态不平衡问题。

## 基于个性化调制器的LoRA大模型微调方法

# 系统实现

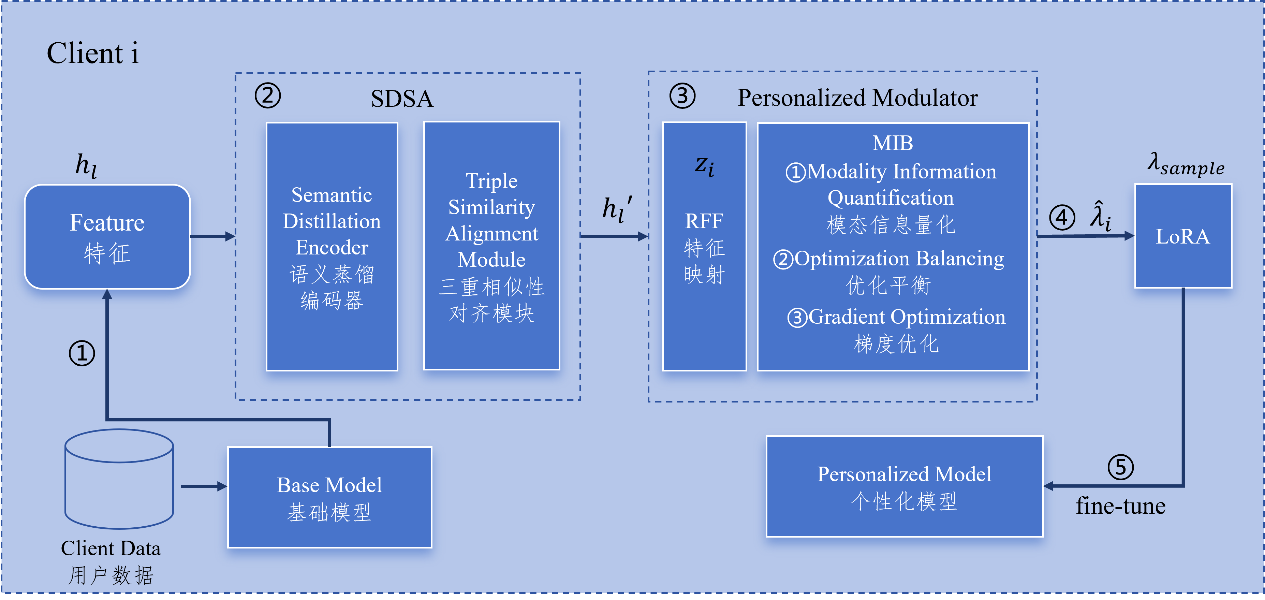


图 4.1 系统框架图

本项目围绕多模态风险内容识别的关键问题，依托SDSA去冗余与语义对齐机制、BMLHF-MIB多模态平衡融合机制与FEDPM联邦双通道学习框架三大核心技术，在多模态数据处理、模型构建、系统部署与优化等方面展开了系统性的工程实现与迭代优化，构建了具有实用价值与工程可落地性的风险内容智能识别系统。

## 数据来源与预处理流程

为验证SDSA去冗余与语义对齐机制，本系统选取两个典型多模态虚假内容数据集作为基础训练与测试资源，分别为中文社交平台微博上的**Weibo 数据集**和英文社交平台Reddit构建的 **Fakeddit 数据集**。在微博数据集中，训练集包含7532条新闻，其中3749条为假新闻，3783条为非假新闻，测试集包含1996篇新闻⽂章。在Fakeddit数据集中，我们从Fakeddit训练集中随机选择30,000对图像-文本对作为我们的训练集，从测试集中随机选择10,000对图像-文本对作为我们的测试集。实验部署在带有PyTorch的NVIDIAGeFor-ce1080TiGPU上。两个数据集的批处理大小都设置为32。

我们的基线模型包括模态特定特征提取器，具有三个完全连接层的编码器，注意网络和分类器。我们将此基线模型称为SDSA\*。SDSA\*+S表示在基线模型上增加三重相似性对齐模块。SDSA\*+I表示用语义蒸馏编码器替换基线模型的编码器。SDSA表示在基线模型中同时添加三重相似性模块和语义蒸馏编码器。我们在表5.1的底部部分报告了结果。

为验证FEDPM联邦双通道学习框架，本研究所使用的实验数据基于 MasakhaNEWS 多语言新闻数据集。该数据集覆盖包括英语（eng）、斯瓦希里语（swa）、豪萨语（hau）、卢干达语（lug）等在内的 16 种非洲地区语言，是当前在多语言预训练与低资源语言研究中广泛采用的标准数据集之一。为符合联邦学习场景的模拟需求，研究中将每个语种视作独立的客户端，构建了符合“非独立同分布（Non-IID）”特征的数据划分结构。

在数据处理方面，首先对原始新闻文本进行标准化预处理，包括去除多余符号、统一编码格式及语种标签清洗。随后采用与预训练模型（如 mBERT 或 XLM-R）配套的分词器进行编码，并依据语种划分数据客户端，严格保证客户端之间数据隔离。所有语种数据均按 8:1:1 的比例划分为训练集、验证集和测试集，并在个性化微调阶段仅使用本地数据进行训练。此外，为模拟真实低资源环境，在数据均衡上适当保留各语种原始分布，以考验模型在低资源场景下的泛化能力与适应性。

## 多模态信息平衡融合机制的实现

在多模态融合模块的实现过程中，项目引入了 MIB（Modal Information Bottleneck）平衡机制，该机制在每一轮训练中对不同模态的特征表达进行动态加权，有效解决了主导模态抑制其他模态表达的问题。通过差异率计算与平衡因子的引入，系统能够动态适配不同模态在融合过程中的权重比例，使得多模态信息得以充分、均衡地利用。其核心处理流程如下伪代码所示：

|  |
| --- |
| 算法2 基于动态权重分配的多模态信息平衡技术 |
| 输入数据：训练集 ，包含文本模态和图像模态以及标签矩阵 ；测试集 ，包含文本模态 和图像模态 。  for each epoch ：   1. 计算各模态特征 ； 2. 计算每个模态的对数几率得分 ； 3. 计算差异率 ； 4. 计算平衡因子 ； 5. 融合多模态特征 ； 6. 通过分类器 得到预测标签 ； 7. 反向传播更新参数；   optimizer.zero\_grad()  L\_total.backward()  更新模型参数；  输出：预测标签。 |

该方法的关键优势在于通过自适应的方式实现特征融合和模型参数更新，提升了模型的鲁棒性和对多模态信息的综合理解能力，为风险内容识别提供了更优的决策依据。

# 测试分析

表 5.1 不同模型在微博和Fakeddit数据集上的比较结果[6]

| Method | Weibo | | | | | | | Fakeddit | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Acc | Rumor | | | Non-rumor | | | Acc | Rumor | | | Non-rumor | | |
| P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 |
| EANN [KDD-2018] | 0.782 | 0.827 | 0.697 | 0.756 | 0.752 | 0.863 | 0.804 | 0.724 | 0.727 | 0.719 | 0.723 | 0.722 | 0.729 | 0.726 |
| SpotFake[BigMM-2019] | 0.892 | 0.902 | 0.964 | 0.932 | 0.847 | 0.656 | 0.739 | 0.819 | 0.801 | 0.848 | 0.824 | 0.839 | 0.790 | 0.813 |
| BDANN[IJCNN-2020] | 0.842 | 0.830 | 0.870 | 0.850 | 0.850 | 0.820 | 0.830 | 0.812 | 0.836 | 0.776 | 0.805 | 0.791 | 0.847 | 0.818 |
| HMCAN [SIGIR-2021] | 0.885 | 0.920 | 0.845 | 0.881 | 0.856 | 0.926 | 0.890 | 0.881 | 0.880 | 0.882 | 0.881 | 0.882 | 0.880 | 0.881 |
| MEAN [SPR-2022] | 0.894 | 0.900 | 0.870 | 0.890 | 0.890 | 0.910 | 0.90 | 0.910 | 0.930 | 0.890 | 0.910 | 0.890 | 0.930 | 0.910 |
| CAFE [WWW-2022] | 0.840 | 0.855 | 0.830 | 0.842 | 0.825 | 0.851 | 0.837 | 0.912 | 0.946 | 0.886 | 0.959 | 0.878 | 0.942 | 0.909 |
| MRML [ICASSP-2023] | 0.897 | 0.898 | 0.887 | 0.892 | 0.896 | 0.905 | 0.901 | 0.840 | 0.819 | 0.874 | 0.846 | 0.865 | 0.807 | 0.835 |
| SDSA\* | 0.898 | 0.916 | 0.888 | 0.902 | 0.880 | 0.909 | 0.894 | 0.941 | 0.946 | 0.936 | 0.941 | 0.936 | 0.947 | 0.942 |
| SDSA\*+S | 0.905 | 0.919 | 0.899 | 0.909 | 0.891 | 0.912 | 0.901 | 0.944 | 0.952 | 0.935 | 0.943 | 0.936 | 0.953 | 0.944 |
| SDSA\*+I | 0.906 | 0.924 | 0.895 | 0.909 | 0.887 | 0.919 | 0.903 | 0.950 | 0.946 | 0.954 | 0.950 | 0.954 | 0.946 | 0.950 |
| SDSA | 0.918 | 0.939 | 0.902 | 0.920 | 0.896 | 0.935 | 0.915 | 0.953 | 0.965 | 0.939 | 0.952 | 0.940 | 0.966 | 0.953 |

从表5.1可以看出，与SDSA\*相比，SDSA\*+S在微博数据集和Fakeddit数据集上分别提⾼了0.8%和0.3%。SDSA\*+I在两个数据集上也分别提⾼了0.9%和0.9%。SDSA\*在两个数据集上分别提⾼了2.2%和1.3%。这些实验结果表明了我们的语义蒸馏编码器和三重相似性对齐模块的有效性。

为验证FEDPM联邦双通道学习框架的研究围绕“已见客户端（seen clients）”情境下的模型个性化能力展开评估，并重点对比了 LoRA、FedL2P 等代表性参数高效微调方法的性能表现。从表5.2中的对比结果可以看出，LoRA 方法在多数语言上取得了较为稳定的准确率表现，如在英语（eng）、皮钦语（pcm）等高资源语言中准确率均在 90% 以上，但在如阿姆哈拉语（amh）等低资源语言上准确率仅为 45.74%。相比之下，FedL2P 方法在整体准确率上略有提升，尤其在 run（81.99%）、som（61.00%）等中等资源语言上相较于 LoRA 有明显进步。

然而，从“wins”列统计的最佳表现次数来看，LoRA 与 FedL2P 在 16 种语言中均未取得明显优势，显示出手工配置的个性化策略仍存在适应性不足的问题，尤其在多语言迁移和极低资源客户端中，个性化方法的差异化效果并不显著。这为后续自动策略生成机制的引入提供了实践动因。

表 5.2 MasakhaNEWS 设置中各语言在已见客户端上的准确率（Mean ± SD）[10]

| **r** | **eng** | **som** | **run** | **fra** | **lin** | **ibo** | **amh** | **hau** | **pcm** | **swa** | **Wins** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **LoRA** | **90.44±0.10** | **60.09±0.32** | **81.37±0.51** | **88.63±0.00** | **83.53±0.54** | **79.83±0.24** | **45.74±0.00** | **75.79±0.00** | **96.05±0.00** | **78.99±0.00** | **0** |
| **FedL2P** | **90.72±0.59** | **61.00±1.16** | **81.99±0.88** | **89.10±0.67** | **83.91±0.00** | **79.66±0.24** | **45.74±0.00** | **76.73±1.12** | **96.05±0.00** | **79.69±0.40** | **0** |

综上，MasakhaNEWS 数据集的多语种分布与数据划分方式充分模拟了真实世界中的非独立、多样性数据环境；而从实验结果来看，现有的参数高效个性化方法虽具一定效果，但在低资源与语言迁移能力上仍存在优化空间。这也进一步说明自动化、稀疏化策略学习机制（如 FedP2EFT）在提升联邦个性化微调能力方面具有现实意义和研究价值。

表 5.3 BMLHF各模块配置在微博和Fakeddit数据集上的性能对比[11]

| **Model** | **Twitter Acc (%)** | **Twitter F1 (%)** | **Weibo Acc (%)** | **Weibo F1 (%)** |
| --- | --- | --- | --- | --- |
| **Base** | 90.12 | 89.95 | 88.29 | 88.15 |
| **Base + HF** | 93.02 | 92.81 | 89.91 | 89.67 |
| **Base + MIB** | 92.37 | 92.20 | 90.58 | 90.35 |
| **BMLHF (Full)** | 94.32 | 94.17 | 91.49 | 91.14 |

从表中结果可以看出，基础模型在 Twitter 和 Weibo 数据集上的准确率分别为 90.12% 与 88.29%，F1 分数分别为 89.95% 与 88.15%，为本研究中最低配置的性能基准。进一步在此基础上加入 HF 模块后，模型性能显著提升，Twitter 上准确率上升至 93.02%，Weibo 上提升至 89.91%，表明层次融合机制在强化模态间协同方面具有良好效果。

而引入 MIB 模块的配置（Base + MIB）同样在两个数据集上表现出稳定增益，准确率分别达到 92.37%（Twitter）与 90.58%（Weibo），说明模态信息平衡机制能够有效缓解“强模态主导”问题，促进弱模态信息的有效利用。

最终，完整模型 BMLHF 在两个数据集上均取得最高准确率和 F1 值，分别为 94.32% / 94.17%（Twitter）和 91.49% / 91.14%（Weibo），显著优于所有消融模型，验证了所提多模态协同机制与模态平衡机制的互补性与协同增益效应。

# 作品总结

## 作品特色与创新点

本项目聚焦于多模态风险内容识别的关键技术难题，结合大模型的深度语义理解能力与跨模态建模优势，系统提出三项核心创新机制，构建了具备强泛化性、高效率、低成本的多模态风险内容识别方案，具有鲜明的技术特色和理论创新。

**1.SDSA数据去冗余与语义对齐机制**：针对多模态数据中普遍存在的冗余信息、噪声干扰与语义不一致问题，提出SDSA机制，从特征层对数据进行选择性清洗与语义对齐，确保输入数据的纯净性与模态协同性，为后续建模提供更具代表性的训练样本，提升识别准确率和模型鲁棒性。

**2.BMLHF多模态平衡融合机制（MIB）**：在多模态融合任务中常见的主模态主导、弱模态退化问题基础上，项目设计并实现BMLHF机制，构建模态平衡约束与混合注意力路径，有效融合不同模态特征，增强模态间的协同表达能力，提升图文与视频风险内容识别的整体性能。

**3.FEDPM联邦多模态双通道学习框架**：为应对数据孤岛和隐私保护等问题，项目构建了联邦式的双通道多模态建模框架FEDPM，结合知识蒸馏与个性化学习策略，显著提升模型在分布异构场景下的跨域泛化能力，实现“数据不出域，模型共进化”的协同训练新范式。

结合BLIP2的轻量级参数微调、基于CLIP的伪造图像构建、三重结构相似性对齐等关键方法，项目在模型构建上兼顾性能与计算效率，形成了统一架构、多模态融合、可扩展性强的技术体系，具有良好的落地潜力与推广前景。

## 应用推广

本项目成果在多模态风险内容检测领域具有广泛的实际应用价值。系统可应用于社交平台虚假新闻识别、诈骗电话与短信检测、伪造音视频监测等多个重点场景，有效提升内容审核、平台治理和舆情管理能力，服务于网络空间治理、安全监管与数字社会建设。

系统架构支持模块化部署和联邦协同更新，可按需适配不同终端与场景，具备从云端平台到边缘设备的灵活应用能力。未来可广泛推广至媒体平台、公安监管、网络安全、舆情监测等领域，为政企机构提供智能化的多模态风险识别解决方案。

## 作品展望

随着多模态生成技术与大模型能力的不断增强，虚假信息的伪造能力持续提升，内容形式更具迷惑性与传播性。本项目在现有成果基础上，后续将在高水平期刊上**发表论文6篇**，**提交发明专利9项**。

同时，我们将探索大模型在低资源场景下的轻量化部署策略与差异化更新机制，推动模型向移动端、边缘设备等轻型终端迁移，助力构建高效、可控、安全的智能化内容识别体系。依托团队现有的研究基础与技术成果，项目将在更大范围内推动多模态大模型技术在实际场景中的应用与落地，为建设清朗健康的数字环境提供有力支撑。

# 参考文献

1. Shukor, M., Couairon, G., & Cord, M. (2022). Efficient Vision-Language Pretraining with Visual Concepts and Hierarchical Alignment (Version 2). arXiv. <https://doi.org/10.48550/ARXIV.2208.13628>
2. Li, J., Li, D., Xiong, C., & Hoi, S. (2022). BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation (Version 2). arXiv. <https://doi.org/10.48550/ARXIV.2201.12086>
3. Li, J., Selvaraju, R. R., Gotmare, A. D., Joty, S., Xiong, C., & Hoi, S. (2021). Align before fuse: Vision and language representation learning with momentum distillation. arXiv preprint arXiv:2107.07651. <https://arxiv.org/abs/2107.07651>
4. Chen, C., Hu, W., Xu, Z., & Zheng, Z. (2021). FedGL: Federated graph learning framework with global self-supervision. arXiv preprint arXiv:2105.03170. <https://arxiv.org/abs/2105.03170>
5. Lao, M., Pu, N., Zhong, Z., Sebe, N., & Lew, M. S. (2023). FedVQA: Personalized federated visual question answering over heterogeneous scenes. In Proceedings of the 31st ACM International Conference on Multimedia (MM '23) (pp. 7796–7807). Association for Computing Machinery. <https://doi.org/10.1145/3581783.3611958>
6. Fei Wu, Shu Chen, Guangwei Gao, Yimu Ji, Xiao-Yuan Jing:Balanced Multi-modal Learning with Hierarchical Fusion for Fake News Detection. Pattern Recognit. 164: 111485 (2025)
7. X. Peng, Y. Wei, A. Deng, D. Wang, D. Hu, Balanced multimodal learning via on-the-fly gradient modulation, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 8238–8247.
8. Y. Wei, R. Feng, Z. Wang, D. Hu, Enhancing multimodal cooperation via samplelevel modality valuation, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 27338–27347.
9. R. Xu, R. Feng, S.-X. Zhang, D. Hu, Mmcosine: Multi-modal cosine loss towards balanced audio-visual fine-grained learning, in: IEEE International Conference on Acoustics, Speech and Signal Processing, 2023, pp. 1–5.
10. Lee, R., Kim, M., Rezk, F., Li, R., Venieris, S. I., & Hospedales, T. (2025). FedP²EFT: Federated learning to personalize parameter efficient fine-tuning for multilingual LLMs. arXiv. <https://arxiv.org/abs/2502.04387>
11. Fei Wu, Shu Chen, Guangwei Gao, Yimu Ji, Xiao-Yuan Jing:

Balanced Multi-modal Learning with Hierarchical Fusion for Fake News Detection. Pattern Recognit. 164: 111485 (2025)