**兴智取——基于深度学习的财政文本信息智能识别抽取模型**

**IMG_256**

**背景信息**

****1. 项目背景与目标****

****1.1 研究背景****

在数字化转型浪潮下，银行业务与外部数据的深度融合变得越来越重要。财政指标数据作为核心外部数据，主要支撑三大业务场景：一是风险评估，可以帮助我们评估地方政府和企业的财政健康状况；二是投资决策，可以通过分析财政收支结构优化资金配置；三是经济预测，可以为银行的发展战略提供数据支持。

金融科技的快速发展对银行的数据处理能力提出了更高的要求。在实际业务中，我们面临着****三大主要挑战****：****首先****是文档格式多样性的问题。财政报告包含图片、PDF、Word等多种格式，文本内容和排版都不统一，给处理带来很大困难。****其次****是人工处理效率低下。传统做法需要人工逐页阅读大量财政报告，往往要花费数小时甚至数天，而且容易出现数据录入错误。最关键的是财政文件本身的专业性很强。报告中不仅包含大量专业术语，数据表达方式也多样化，既有具体数值，也有增长率、占比等不同形式。传统的文本解析技术往往难以准确理解和提取这些信息。

****1.2 项目目标****

本项目针对银行财政文件信息提取的痛点，本项目主要设定了****三个核心目标****：第一，实现多格式文件的统一解析。我们的系统能够同时处理PDF、Excel表格、Word文档等多种格式的财政文件，将它们转换为统一的数据结构。第二，实现关键财政指标的智能识别。系统不仅能够准确定位数值信息，还能理解其时间属性、空间属性以及计量单位等完整语义信息。第三，构建多策略融合的智能提取框架。通过规则策略和深度学习模型的结合，配合大语言模型处理歧义情况，全面提升复杂场景下的识别准确率。通过这三个目标的实现，进而推动财政报告指标提取工作的全面提质增效。

拟构建一套基于深度学习的智能信息提取模型，旨在实现跨文件格式、多场景的财政指标精准识别。与传统方法相比，本模型具有显著的技术创新价值：首先，通过****多类型文件解析技术****，可有效处理PDF、Excel、Word等多种格式文件，解决传统单一文件处理的局限性；其次，****创新性地融合规则匹配、最长公共子串算法、大模型匹配方法****，显著提升信息提取的准确率和召回率；再者，****引入大模型辅助提取和动态置信度评估机制****，进一步增强模型的鲁棒性和泛化能力。这一技术方案不仅能显著降低银行财政文件信息处理的人力成本，还将为金融机构的区域经济研究、信贷风险评估提供更加精准、及时的数据支撑。

**IMG_257**

**核心模式**

****2. 系统架构设计****

****2.1 设计原则****

本系统遵循****“分层架构、模块解耦、策略融合、效果优先”****的设计理念。采用三层架构设计，实现数据处理、模型分类和信息抽取的功能解耦，确保系统的可扩展性和维护性。在具体实现中，采用多策略融合的方式提升信息抽取的准确性和泛化能力，通过规则策略保证高置信度场景的精确匹配，通过深度学习模型提升复杂场景的识别能力，通过大语言模型补充处理歧义和变体表达。同时，系统设计注重计算效率与精度，采用多级过滤和批处理机制，实现了大规模文档的高效处理。

****2.2 技术路线****

系统采用****“解析-分类-抽取”****的三段式技术路线。首先，通过文档标准化预处理和OCR等技术，将多源异构的文档数据转换为规范化文本；其次，基于BERT预训练模型构建二分类器，对文本进行初筛，识别包含目标指标和数值信息的候选句；最后，采用多策略融合的方式进行精确信息抽取，结合关键词精确匹配、最长公共子串匹配和大语言模型推理等方法，实现对财政预算指标及其对应数值的精确识别和提取。在处理流程中，系统通过分层设计和多轮迭代，不断优化提取结果，确保抽取的准确性和完整性。

****2.3 系统架构****

系统采用****三层架构设计****，实现了从原始文档到目标信息的完整处理链路。数据解析层负责多源文档的标准化处理和文本预处理，为上层提供高质量的文本输入；分类模型层基于深度学习技术，实现对包含目标信息的文本的智能识别，显著提升了处理效率；智能抽取层通过多策略融合方式，实现对目标指标及其数值的精确提取。三层之间通过标准化接口进行数据传递，既保证了各层的独立性，又实现了功能的有机整合，形成了一个完整的信息抽取处理体系。

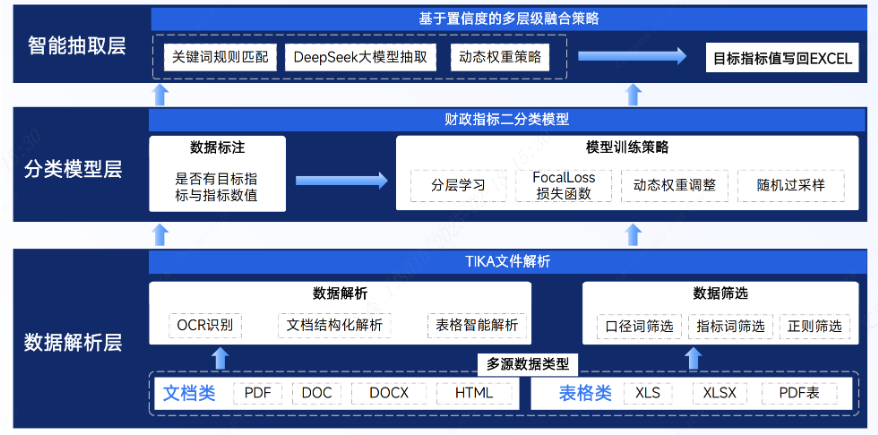
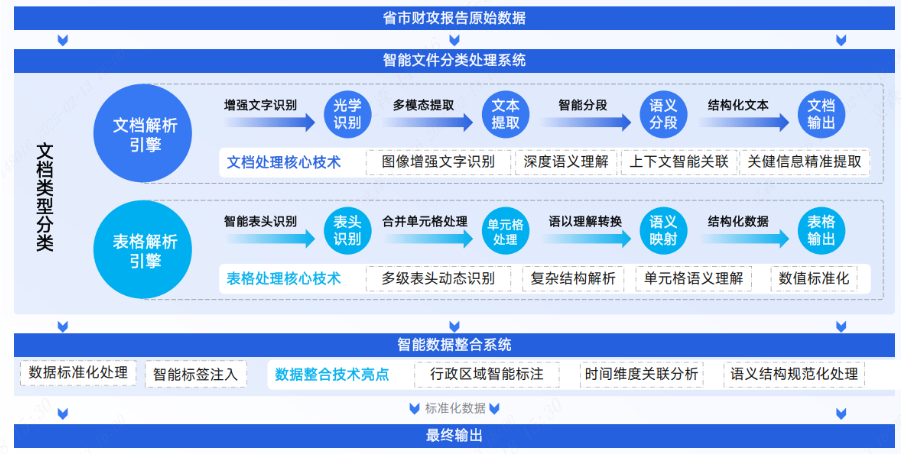


图1 技术框架图

****2.3.1 数据解析层****

数据解析层主要解决多源异构数据的统一处理问题。该层首先对输入的财政文档进行格式识别和分类，借助强大文件解析工具****TIKA开源库以及自构的各种类型文件解析器****建立标准化的文档处理流程，确保不同格式文档都能被正确处理。对于PDF等图像化文档，通过OCR技术进行文字识别，实现文本信息的准确提取；对于Word和Excel等结构化文档，则通过特定的解析器提取文本内容。在文本预处理阶段，实现了多级的文本清洗功能，包括特殊字符处理、编码统一、冗余信息过滤等。同时，针对财政文档的特点，设计了专门的分句算法，既保证了语义的完整性，又提高了后续处理的精确度。该层还实现了文档目录结构的优化功能，通过建立统一的存储体系，提高了数据访问效率。此外，通过实现批处理机制，显著提升了大规模文档的处理效率，为上层提供高质量的文本数据支持。

图2 数据解析层

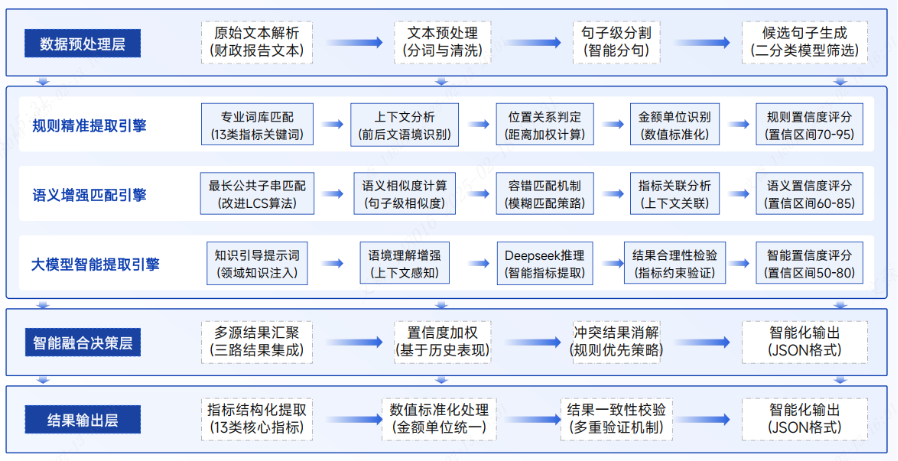
****2.3.2 分类模型层****

分类模型层主要负责对文本进行智能分类，识别包含目标财政指标和数值信息的文本片段。该层****基于BERT预训练模型构建了专门的二分类器****，通过对大量标注数据的学习，实现了对目标文本的准确识别。在模型设计上，采用了特殊的网络结构，通过冻结部分底层参数，保留语言理解能力的同时，提升了模型在特定任务上的表现。训练过程中，通过精心设计的数据增强策略，扩充了训练样本的多样性，提升了模型的泛化能力。同时，采用了针对性的损失函数设计和优化策略，有效解决了样本不平衡问题。模型还实现了基于注意力机制的文本特征提取，能够准确捕捉文本中与财政指标相关的关键信息。通过设置合理的置信度阈值，实现了对候选句的精确筛选，大幅提升了后续信息抽取的效率和准确率。

图3 分类模型层

****2.3.3 智能抽取层****

智能抽取层承担着从候选文本中精确提取目标指标和对应数值的关键任务。该层采用了创新的多策略融合框架，综合运用多种抽取方法，以应对不同场景下的信息提取需求。在处理流程上，首先通过关键词精确匹配策略处理标准表述的情况，确保高置信度场景下的提取准确性；对于表述变体和同义表达，则通过最长公共子串匹配算法进行处理，提升了系统的容错能力。在处理复杂语义和特殊表达时，通过调用大语言模型进行辅助理解和提取，有效提升了系统的智能化水平。该层还设计了基于多维度特征的置信度评分机制，通过对不同策略结果的智能融合，实现了最优提取结果的选择。在提取过程中，采用多轮迭代的方式，优先处理高置信度的结果，然后逐步扩大处理范围，最终实现了对目标信息的全面准确提取。

图4 智能抽取层

****2.4 核心功能流程****

本系统首先通过数据解析层对输入的多源文档进行统一处理，将PDF、Word、Excel等不同格式的文档转换为标准化文本，并进行基础的文本清洗和规范化处理。文本经过分句处理后，由BERT二分类模型进行初筛，识别出包含目标指标和数值信息的候选句。对于识别出的候选句，系统采用两轮提取策略：首轮主要处理报告类文档，优先使用高置信度的关键词匹配方法提取指标信息；对于未能成功提取的指标，进入第二轮处理，扩大到非报告类文档，综合使用LCS匹配和大语言模型推理等方法进行补充提取。最后，系统对提取结果进行验证和整合，确保输出数据的准确性和完整性，并按照目标模板格式进行结果输出。

****3. 数据解析层设计与实现****

****3.1 多源数据解析模块****

多源数据解析模块是整个系统的数据获取基础，主要负责处理包括PDF、Word、Excel和HTML在内的多种格式文档，将非结构化和半结构化的财政预算文档转换为标准化文本格式。针对PDF文档，系统采用OCR技术进行文字识别，通过版面分析算法保持文档的逻辑结构，同时优化图表区域的文本提取策略。对于Word文档，重点关注文本内容的完整提取，通过编码统一确保中文字符的正确解析。在处理Excel文档时，采用特定的表格解析策略，准确识别单元格间的关系，保留数值与指标的对应关系。HTML文档则通过DOM结构解析，精准提取目标文本内容，并进行格式化处理。模块设计了统一的数据解析接口，实现了不同格式文档的标准化处理，为后续的文本分析和信息提取奠定基础。

****3.2 数据处理筛选模块****

数据处理筛选模块主要通过folder\_preprocess实现对解析后的文件进行初步处理和筛选。在文档预处理阶段，系统仅保留pdf、html、xlsx、xls、doc、docx格式的文件，对zip格式的文件进行解压并处理，删除空文件，并将所有文件扩展名统一转换为小写。在文本规范化处理环节，对文件名中包含“报告”的xlsx和xls文件进行过滤，避免重复内容。系统会将解压后的文件进行标准化组织，删除空文件夹，建立规范的存储结构。

对于文本内容的处理，系统通过report\_read和table\_read模块分别处理报告类文件和表格类文件。在报告类文件处理中，使用基本的句子切分规则（如句号、分号、感叹号、问号）进行分句，并筛选出包含“元”字或与目标指标高度相关元素的句子。同时，对包含“预算”、“决算”或年份信息的句子进行保留，确保关键信息不丢失。对于表格类文件，则通过处理合并单元格、标准化表头等方式进行数据规范化。

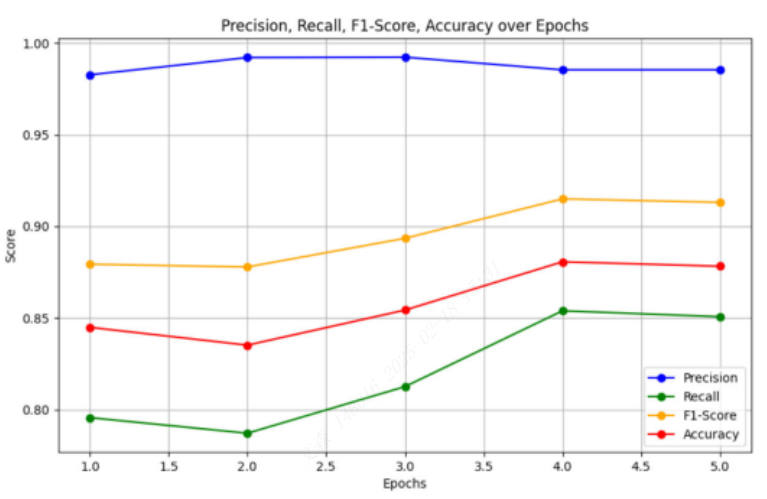
****4. 模型分类层设计与实现****

****4.1 BERT二分类模型****

系统基于BERT-base-Chinese预训练模型构建了面向财政预算文本的二分类模型。模型架构上采用了典型的预训练语言模型微调范式，在BERT的[CLS]标记输出上连接了一个Dropout层（dropout率为0.1）和一个二分类的全连接层。为了提高模型的训练效率和泛化能力，采用了分层学习率策略：对BERT的前8层参数进行冻结，仅微调后4层参数，这种设计既保留了BERT底层获取的通用语言特征，又允许高层特征适应具体任务。在模型训练方面，采用了FocalLoss损失函数来解决样本不平衡问题，并引入了基于class\_weight的动态权重调整机制。训练数据采用了随机过采样（RandomOverSampler）策略来平衡正负样本比例，并通过TensorDataset实现了高效的数据加载和批处理。

****4.2 模型训练策略****

模型训练过程采用了一系列优化策略以提升模型性能。首先，在优化器选择上使用AdamW，配合线性学习率衰减策略，初始学习率设为2e-5，并在总训练步数的前10%实现了预热（warm-up）。为防止过拟合，实现了基于验证集性能的早停机制，当验证集性能连续3个epoch没有提升时终止训练。在数据处理方面，将最大序列长度设置为256 tokens，采用batch\_size为16的小批量训练策略，这些参数的选择是基于实验优化的结果。训练过程中通过TensorBoard记录训练损失、验证集性能等指标，实现了训练过程的可视化监控。模型评估采用了多个指标，包括准确率、ROC-AUC、精确率和召回率等，最终测试集上取得了****高于0.9的ROC-AUC score****，体现了模型在实际应用中的良好表现。

图2 分类模型训练效果图

****5. 智能抽取层设计与实现****

****5.1 多策略抽取框架****

多策略抽取框架构建了一个****三层抽取体系****，通过不同策略的优势互补提升系统性能。在****关键词精确匹配策略****中，系统基于完全字符匹配原则构建了包含[“一般公共预算收入”, “税收收入”]等13个目标指标的匹配模式，采用正则表达式对数值金额进行标准化处理，同时通过年份标识判断、前缀匹配和距离计算等机制提升匹配准确性。对于未能通过精确匹配识别的内容，系统引入了****基于动态规划的LCS最长公共子串匹配算法****，通过计算文本片段与目标指标的重合度，实现了对指标变体表述的识别。

为了处理更复杂的语义理解场景，系统集成了****基于星火大模型的智能匹配策略****。该策略通过构造专门的prompt模板，引导大模型从目标文本中精准识别财政指标及其对应数值，并通过JSON格式规范化输出结果。系统还实现了基于批处理的并发调用机制，设置了合理的速率限制，并构建了完整的异常处理流程。三种策略通过统一的接口规范实现了数据的互通，形成了一个完整的抽取框架。

****5.2 置信度评分体系****

置信度评分体系设计了一套****完整的多维度评分标准****。对于关键词精确匹配策略，系统从基础匹配、语境特征和数值特征三个维度进行评分：基础匹配包括指标词完全匹配（75分基准）和前缀匹配（5分加成）；语境特征包括年份标识（10分加成）和指标词与数值的距离评估（最多10分扣除）；数值特征包括金额合理性判断（5分加成）和单位标准化评估。通过这些维度的综合评分，为每个匹配结果生成一个范围在60-100分之间的置信度得分。

对于LCS匹配策略，系统设计了基于匹配度的评分机制。主要考虑最长公共子串占比、文本片段距离、数值位置关系等因素，通过加权计算得出置信度分数。对于大语言模型的输出结果，系统基于模型的输出概率和对输出格式的规范性进行评分。同时，系统还实现了跨策略的置信度归一化处理，确保不同策略的得分具有可比性。这种多维度的评分体系为结果筛选和融合提供了可靠的依据。

对于大语言模型匹配策略，系统设计了四个置信度等级：90-100分表示指标完全匹配且上下文明确无歧义；80-90分适用于指标完全匹配但数值位置较远的情况；60-79分用于基本匹配但存在多个候选值的情况；60分以下表示存在较大歧义。评分还考虑了数值的合理性、单位一致性等因素，并通过JSON格式规范化输出结果。三种策略的评分结果通过归一化处理实现了跨策略的可比性，为最终的结果筛选和融合提供了可靠依据。

****5.3 多轮迭代策略****

系统实现了两轮迭代的智能抽取策略。第一轮处理以报告类文件为主，设置0.7的较高置信度阈值，优先使用keyword\_match方法进行高精度抽取。对于每个待提取指标，首先进行年份和前缀标识识别，然后通过pattern模式匹配提取数值信息，最后进行单位标准化处理。第二轮处理扩展到非报告类文件，降低置信度阈值至0.5，综合使用三种抽取策略。系统按照"先易后难、先高置信后低置信"的原则进行迭代处理，并在结果选择时实现了基于置信度的优先级策略，有效平衡了准确率和召回率的需求。通过这种分层迭代的机制，系统最大化利用了不同文档类型和不同抽取策略的优势，提升了整体的抽取性能。

**IMG_256**

**创新亮点**

本系统的创新性主要体现在以下****三个方面****。

首先，项目创新性地将****BERT二分类模型作为初筛机制****，通过冻结模型前8层参数的方式实现了高效的迁移学习，显著提升了模型在财政领域文本上的适应性。同时，系统设计了基于置信度加权的多策略融合机制，将关键词精确匹配、最长公共子串匹配和大模型辅助提取有机结合，不仅提升了系统的鲁棒性，也为不同场景下的文本提取提供了灵活的解决方案。

在架构设计上，系统实现了****数据解析层、分类模型层、智能抽取层****的三层架构，通过标准化接口实现了模块间的低耦合高内聚。特别是在智能抽取层，创新性地设计了多维度的置信度评分体系，实现了不同策略结果的可靠融合。系统还建立了完整的异常处理和日志记录机制，通过批处理和并行处理优化提升了系统效率，体现了一定的工程实践价值。

此外，本项目还制定了****动态置信度评分体系****，通过多维度特征的综合评估，实现了跨策略结果的可靠融合。同时，分层迭代的处理策略也很好地平衡了准确率和召回率的需求。