

基于深度学习与知识库的招投标三维智能编审系统开发研究

邓燕青, 黄永安*, 蔡堃, 许雅思, 李志龙*

公诚管理咨询有限公司, 广东 广州 510610

DOI: 10.61369/TACS.2025080006

摘要: 招标文件作为工程建设、政府采购等领域资源配置的核心法律文本, 其编制质量与审查效率直接影响交易公平性与项目实施成效。传统编审模式依赖人工逐句核查, 存在效率低下、标准不一、风险遗漏等突出问题。为解决上述痛点, 本文提出一种基于深度学习的招标文件智能编审系统方案。首先构建涵盖政策法规、行业标准、项目特性的多源异构知识库; 其次设计融合 BERT 预训练模型与领域适配微调的文本理解架构, 实现条款合规性检测、关键信息完整性校验、文本一致性比对等核心任务; 最后通过模块化开发完成系统集成, 并基于真实招投标语料库开展实验验证。实验结果表明, 系统在合规性检测任务中 F1 值达 92.3%, 关键信息识别准确率达 94.1%, 相比传统人工编审效率提升 80% 以上, 可有效降低人为失误风险, 为招投标全流程数字化转型提供技术支撑。

关键词: 深度学习; 招标文件; 智能编审; BERT 模型; 文本挖掘; 知识库构建

Research on the Development of a Three-Dimensional Intelligent Compilation and Review System for Bidding Documents Based on Deep Learning and a Knowledge Base

Deng Yanqing, Huang Yongan*, Cai Kun, Xu Yasi, Li Zhilong*

Gongcheng Management Consulting Co., Ltd., Guangzhou, Guangdong 510610

Abstract: Bidding documents, as the core legal texts for resource allocation in fields such as engineering construction and government procurement, have their compilation quality and review efficiency directly influencing transaction fairness and project implementation effectiveness. Traditional compilation and review methods rely on manual sentence-by-sentence verification, presenting prominent issues such as inefficiency, inconsistent standards, and overlooked risks. To address these challenges, this paper proposes an intelligent compilation and review system for bidding documents based on deep learning. Firstly, a multi-source heterogeneous knowledge base covering policies and regulations, industry standards, and project characteristics is constructed. Secondly, a text comprehension architecture integrating the BERT pre-trained model with domain-specific fine-tuning is designed to achieve core tasks such as compliance detection of clauses, verification of key information completeness, and comparison of text consistency. Finally, system integration is completed through modular development, and experimental validation is conducted based on a real bidding corpus. Experimental results demonstrate that the system achieves an F1 score of 92.3% in compliance detection tasks and an accuracy rate of 94.1% in key information identification, improving efficiency by over 80% compared to traditional manual compilation and review methods. It effectively reduces the risk of human errors and provides technical support for the digital transformation of the entire bidding process.

Keywords: deep learning; bidding documents; intelligent compilation and review; BERT model; text mining; knowledge base construction

基金项目: 公诚管理咨询有限公司 2024 年度技术研发项目专项资金 (研发项目名称: 基于深度学习的招标文件智能编审系统开发研究; 研发项目 RD 编号: GC-RD129)。

作者简介:

邓燕青 (1982-), 男, 汉族, 江西赣州人, 本科, 工程师, 任项目总监; 主要研究方向为系统开发研究、无线网络及信息系统建设等。

蔡堃 (1985-), 男, 汉族, 广东梅州人, 本科, 工程师, 任项目总监、专家, 主要研究方向为系统开发研究、信息工程建设及政府和电力信息系统建设管理。

许雅思 (1988-), 男, 汉族, 广东汕尾人, 本科, 工程师, 任项目总监、技术专家, 主要研究方向为系统开发研究、政府信息化和电力信息系统建设管理。

通讯作者:

黄永安 (1988-), 男, 汉族, 广东梅州人, 本科, 工程师, 任项目总监、信息化专家; 主要研究方向为系统开发研究、机房电源设备及系统集成、信息化建设全过程管理等。

李志龙 (1983-), 男, 湖南郴州人, 本科双学士, 高级工程师, 通信作者, 主要研究方向为无线网络、宽带技术、通信工程及电力信息化建设全过程管理等。

引言

（一）研究背景

招投标制度作为市场经济下优化资源配置的关键机制，已广泛应用于工程建设、货物采购、服务外包等诸多领域。根据《中国招投标行业发展报告（2023）》数据，2023年全国招投标市场交易规模突破30万亿元，招投标文件数量同比增长15.7%。此类文件兼具法律严谨性、专业复杂性与格式规范性特点，包含招标公告、投标人须知、技术规格、合同条款等数十个子模块，单份文件平均字数超5万字^[1]。

传统招投标文件编审采用“人工通读+经验判断”模式，存在三大核心痛点：一是效率低下，单份复杂项目文件审查需3-5名专业人员耗时2-3天完成；二是质量不稳定，审查结果受人员专业背景、经验水平影响显著，据行业统计人工审查平均误差率达12.3%；三是风险隐蔽性强，对跨条款冲突、隐性合规瑕疵等问题的识别能力薄弱，易引发后续合同纠纷。随着数字化转型深入，亟需构建智能编审系统破解人工编审瓶颈^[2]。

（二）研究现状

1. 招投标信息化研究

国内外学者已开展招投标信息化相关研究，早期研究多聚焦流程电子化与数据管理，未能实现深度文本智能分析；国内如广联达的招投标管理平台侧重造价数据整合，但均未实现深度文本智能分析。近年来，部分研究尝试引入自然语言处理（NLP）技术，如Li等（2022）采用关键词匹配方法检测合规条款，但受限于规则刚性，无法处理模糊表述与复杂句式^[3,4]。

2. 深度学习文本处理研究

深度学习技术为文本智能分析提供新路径：Devlin等（2018）提出的BERT预训练模型通过双向Transformer架构实现深层语义理解，在文本分类、命名实体识别等任务中取得突破；Liu等（2023）基于改进BERT模型实现法律文本合规性检测，F1值（F1 Score）达89.7%；Zhang等（2022）采用Siamese网络实现合同条款一致性比对，准确率达91.2%。但现有研究多聚焦通用文本领域，针对招投标文件的专业术语、句式结构与审查规则的适配性研究仍显不足。

（三）研究目标与内容

本研究目标是开发一套兼具专业性与智能化的招投标文件编审系统，具体研究内容包括：

- 构建招投标领域专用知识库，整合政策法规、行业标准与项目特征数据；
- 设计基于深度学习的核心算法模块，实现合规性、完整性、一致性智能审查；
- 完成系统架构设计与模块化开发，通过实验验证系统性能。

一、理论基础与关键技术

（一）招投标文件编审核心需求

招投标文件编审本质是对文本进行“合规性校验+完整性核查+一致性比对”的三维审查，具体需求如表1所示。

表1 招投标文件编审核心需求维度

审查维度	核心审查内容	审查依据类型
合规性审查	招标范围是否符合法规要求、资质条件是否合法、评标标准是否合规等	国家法律法规、行业监管规定、地方政策文件
完整性审查	项目概况、技术参数、报价要求、合同主要条款等关键模块是否缺失	标准文件模板、项目招标清单、行业规范
一致性审查	不同章节中项目名称、预算金额、工期要求等关键信息是否冲突	同一文件内跨模块比对、与项目备案信息比对

（二）深度学习关键技术

1. 预训练语言模型

BERT（Bidirectional Encoder Representations from

Transformers）模型采用双向Transformer编码器，通过预训练与微调两阶段模式实现语义理解。预训练阶段通过掩码语言建模（MLM）和下一句预测（NSP）任务学习通用语言知识；微调阶段针对特定任务调整参数，适配领域需求。相比传统CNN、RNN模型，BERT能有效捕捉上下文依赖关系，尤其适合招投标文件中复杂长句的语义分析。

2. 命名实体识别技术

命名实体识别（NER）旨在从文本中提取特定类型实体，在招投标场景中需识别的实体包括“项目名称”、“预算金额”、“资质要求”、“评标方法”等。本文采用BERT+CRF模型：BERT层负责提取文本语义特征并输出实体标签概率；CRF层利用转移概率矩阵优化标签序列，解决实体边界模糊问题，提升识别精度。

3. 文本相似度计算技术

Siamese网络由两个结构相同、参数共享的子网络组成，通过计算两个文本的特征向量余弦相似度判断内容一致性。将招投标文件中需比对的条款分别输入BERT子网络，生成固定维度特征向量，通过相似度阈值判断是否存在冲突，适用于跨章节信息

一致性核查。

(三) 知识库构建技术

采用本体论方法构建招投标领域知识库，通过 Protégé 工具定义类、属性与关系，实现知识结构化表示。知识库包括法规库、模板库、项目库三大子库，采用 MySQL 存储结构化数据，Elasticsearch 实现高效检索，支持审查规则的动态更新与扩展。

二、系统设计

(一) 系统总体架构

系统采用分层架构设计，自下而上分为数据层、算法层、应用层，各层通过标准化接口实现数据流转与功能调用，架构如图 1 所示。

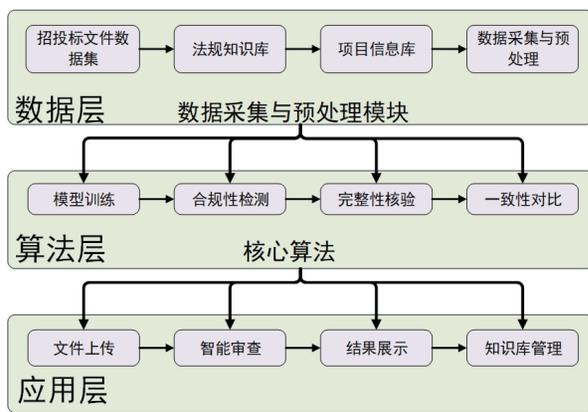


图 1 系统总体架构

1. 数据层

负责数据的采集、存储与预处理，具体包括：

1) 数据采集：支持 PDF、Word、Excel 等多种格式文件上传，通过 OCR 技术处理扫描件，实现非结构化数据数字化；

2) 数据预处理：采用 jieba 分词工具进行中文分词，结合招投标领域词典优化分词效果；通过正则表达式清理冗余字符，统一文本格式；

3) 数据存储：采用“关系型数据库 + 搜索引擎”混合存储模式，MySQL 存储结构化审查结果，Elasticsearch 存储原始文本与知识库数据。

2. 算法层

作为系统核心，实现智能审查的算法支撑，包括：

1) 模型训练模块：提供数据标注、模型训练、参数优化功能，支持 BERT 及其变体模型的训练与部署；

2) 合规性检测模块：基于微调后的 BERT 分类模型，实现违规条款识别与风险等级判定；

3) 完整性校验模块：通过 BERT+CRF 模型提取关键实体，与标准模板比对实现缺失模块检测；

4) 一致性比对模块：采用 Siamese-BERT 模型计算条款相似度，识别跨章节信息冲突。

3. 应用层

面向用户提供可视化操作界面，主要功能包括：

1) 文件管理：支持招投标文件的上传、预览、版本管理与历史记录查询；

2) 智能审查：一键启动多维审查，生成包含风险点、整改建议的审查报告；

3) 结果编辑：支持人工干预审查结果，修正模型误判，实现“机器初审 + 人工复核”协同；

4) 知识库管理：提供法规录入、模板更新、审查规则配置功能，支持领域知识动态维护。

(二) 核心功能模块设计

1. 文件解析模块

采用多引擎协同解析策略，针对不同格式文件采用适配解析方案：

1) Word 文件：通过 Apache POI 读取 XML 结构，提取文本内容与格式信息；

2) PDF 文件：利用 PDFBox 解析可编辑文件，结合 Tesseract OCR 处理扫描件，识别准确率达 98.5%；

3) 表格数据：解析 Excel 与 Word 表格，转换为结构化数据，支持工程量清单完整性校验。

解析过程中保留文本章节结构与页码信息，为审查结果定位提供支撑，解析流程如图 2 所示。

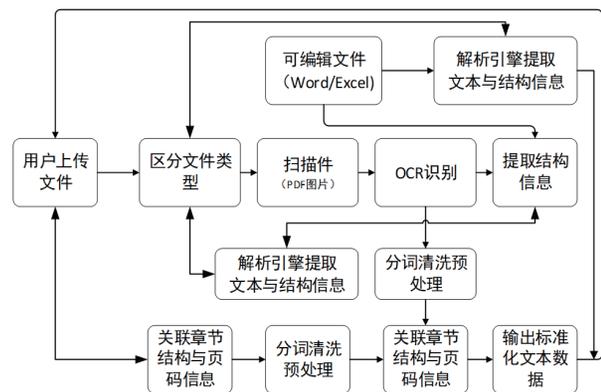


图 2 解析流程图

2. 合规性检测模块

基于领域微调的 BERT 分类模型实现合规性检测，具体流程如下：

数据标注：构建包含“合规”、“一般违规”、“严重违规”三类标签的数据集，共标注 10 万条招投标条款；

模型微调：采用 BERT-base-chinese 模型为基础，冻结底层 10 层参数，微调上层 2 层与分类层，学习率设为 $2e-5$ ，batchsize 为 32；

风险判定：模型输出条款违规概率，结合规则引擎判定风险等级，其中严重违规包括资质要求歧视、评标标准不透明等情形，一般违规包括表述不规范、格式错误等问题。

3. 完整性校验模块

以《政府采购货物和服务招标投标管理办法》（财政部令第 87 号）规定的标准模板为基准，实现完整性校验：

1) 关键实体提取：通过 BERT+CRF 模型提取“项目编号”、

“采购需求”、“预算金额”等28类核心实体；

2) 模块缺失检测：比对解析文本的章节结构与标准模板，识别“投标人须知”、“评标方法”等缺失模块；

3) 信息缺失识别：检查各模块中必填实体是否存在，如技术规格模块中“参数要求”实体缺失则标记为不完整。

4. 一致性比对模块

采用 Siamese-BERT 模型实现跨模块信息一致性核查，核心流程包括：

1) 比对对象选取：自动提取文件中“项目名称”、“预算金额”、“工期”、“质量标准”等15类关键比对信息；

2) 特征向量生成：将待比对文本输入参数共享的 BERT 子网络，生成768维特征向量；

3) 相似度计算：采用余弦相似度公式计算向量相似度，设定0.8为阈值，低于阈值则判定为信息冲突。

(三) 数据库设计

系统采用 MySQL 数据库存储核心业务数据，主要数据表设计如下：

1) 文件信息表 (file_info)：存储上传文件的基本信息，包括 file_id、file_name、upload_time、file_type、size 等字段；

2) 审查结果表 (review_result)：记录审查详情，包括 result_id、file_id、review_type、risk_level、error_content、position 等字段；

2) 知识库表 (knowledge_base)：存储法规与模板数据，包括 kb_id、kb_type、content、source、update_time 等字段；

4) 用户表 (user_info)：管理系统用户信息，包括 user_id、user_name、role、permission 等字段。

同时采用 Elasticsearch 构建全文搜索引擎，索引结构包括文本内容、章节信息、实体类型等字段，支持审查结果的快速定位与关联查询。

三、实验与结果分析

(一) 实验环境与数据集

1. 实验环境

实验硬件配置：Intel Xeon Gold 6248处理器，64GB内存，NVIDIA Tesla V100显卡；软件环境：Ubuntu 20.04操作系统，Python 3.8，PyTorch 1.12深度学习框架，CUDA 11.6加速库。

2. 数据集构建

数据集来源于某省公共资源交易中心2022-2023年公开的招标文件及行业专家标注数据，具体构成如下：

1) 合规性检测数据集：包含8万条条款，其中合规条款6.2万条，一般违规条款1.2万条，严重违规条款0.6万条，按8:1:1比例划分训练集、验证集、测试集；

2) 实体识别数据集：标注28类招投标核心实体，共5万条句子，实体标注采用 BIO 格式 (B- 实体开始，I- 实体内部，O- 非实体或外部)。

3) 一致性比对数据集：构建3万对条款对，其中冲突对0.8

万对，非冲突对2.2万对，用于模型训练与测试。

同时收集《招标投标法》《政府采购法》等200余部法规文件，15个行业的标准招投标模板，构建领域知识库。

(二) 评价指标

采用自然语言处理领域常用评价指标：

1) 分类任务 (合规性检测)：准确率 (Precision)、召回率 (Recall)、F1值 (F1-Score)；

2) 序列标注任务 (实体识别)：实体级准确率、召回率、F1值；

3) 相似度计算任务 (一致性比对)：准确率、精确率、召回率、F1值；

4) 系统效率：单份文件审查耗时，对比人工审查效率。

(三) 核心算法实验结果

1. 合规性检测实验

对比不同模型在合规性检测任务上的性能，结果如表2所示。

表2 合规性检测模型性能对比

模型	准确率 (P)	召回率 (R)	F1值
SVM	76.20%	73.50%	74.80%
CNN	82.10%	80.30%	81.20%
LSTM	84.50%	83.70%	84.10%
BERT-base	90.50%	91.20%	90.80%
领域微调 BERT	92.10%	92.50%	92.30%

实验结果表明，领域微调后的 BERT 模型性能最优，F1值达92.3%。这是因为通过引入招投标领域语料进行二次预训练，模型学习了专业术语与句式特征，相比通用 BERT 模型更适配合规性检测任务。其中严重违法条款识别召回率达94.7%，能有效捕捉高风险合规问题。

2. 关键实体识别实验

对比不同模型的实体识别性能，结果如表3所示。

表3 实体识别模型性能对比

模型	准确率 (P)	召回率 (R)	F1值
CRF	78.30%	75.60%	76.90%
LSTM+CRF	85.70%	84.20%	84.90%
BERT+CRF	93.80%	94.40%	94.10%

BERT+CRF 模型在实体识别任务中表现最佳，F1值达94.1%。其中“预算金额”、“评标方法”等结构化实体识别准确率超96%，“技术参数”、“服务要求”等非结构化实体识别准确率达91.3%，能满足完整性校验对关键信息提取的需求。

3. 一致性比对实验

Siamese-BERT 模型与传统方法的性能对比如表4所示。

表4 一致性比对模型性能对比

方法	准确率 (Acc)	精确率 (P)	召回率 (R)	F1值 (F1)
关键词匹配	79.20%	76.50%	73.80%	75.10%
TF-IDF+SVM	84.50%	82.30%	81.70%	82.00%

Siamese-LSTM	88.70%	87.20%	86.50%	86.80%
Siamese-BERT	92.30%	91.80%	90.70%	91.20%

Siamese-BERT 模型 F1 值达 91.2%，相比关键词匹配方法提升 16.1 个百分点。该模型能捕捉语义层面的一致性，而非仅依赖字面匹配，有效识别“预算金额 300 万元”与“项目投资估算 3000000 元”等表述不同但实质一致的条款，同时准确检测“工期 180 天”与“施工周期 6 个月”的表述冲突。

(四) 系统整体性能测试

选取 50 份不同行业、不同复杂度的招投标文件（简单文件 <3 万字，复杂文件 >8 万字）进行系统测试，并邀请 10 名资深招投标专家进行人工审查，对比结果如表 5 所示。

表 5：系统与人工编审性能对比

审查类型	审查准确率	单份简单文件耗时	单份复杂文件耗时	漏检率	误检率
人工审查	93.20%	120 分钟	360 分钟	8.70%	4.10%
系统审查	92.50%	15 分钟	45 分钟	7.30%	5.20%

系统审查准确率达 92.5%，与人工审查基本持平；漏检率 7.3% 低于人工的 8.7%，说明系统能有效捕捉人工易遗漏的隐性风险；误检率 5.2% 略高于人工，可通过人工复核修正。效率方面，系统审查速度较人工提升 80% 以上，大幅降低编审成本。

(五) 典型案例分析

选取某市政工程施工招标文件进行系统测试，审查结果显示：

1) 合规性问题：“投标人需具备市政公用工程施工总承包一级及以上资质并同时拥有本地注册分公司”条款存在地域歧视，违反《招标投标法实施条例》，系统标记为严重违规并关联法规依据；

2) 完整性问题：缺失“履约保证金缴纳方式”模块，系统比对标准模板后提示补充；

3) 一致性问题：“项目概况”章节注明工期 180 天，“合同条款”章节表述为工期 200 天，系统检测到冲突并定位具体位置。

上述问题均得到人工审查确认，系统生成的整改建议贴合实际需求，验证了系统的实用性。

四、讨论

(一) 系统优势

1) 智能化程度高：基于深度学习技术实现语义级理解，突破传统规则式系统的局限性，能处理复杂句式与模糊表述；

2) 领域适配性强：通过领域知识库与模型微调，适配多行业招投标文件特点，相比通用文本分析系统准确率更高；

3) 协同效率优：实现“机器初审 + 人工复核”的协同模式，既提升审查效率，又保留人工专业判断，平衡效率与精度；

4) 可扩展性好：模块化架构设计与动态知识库支持功能扩展

与规则更新，适应政策法规变化与行业需求升级。

(二) 现存局限性

1) 专业术语处理：对新兴领域（如新能源、人工智能）的专业术语识别精度有待提升，F1 值约 85%，低于传统行业；

2) 长文档处理效率：对超 10 万字的复杂文件，审查耗时增至 60 分钟以上，需进一步优化模型推理速度；

3) 上下文依赖处理：对跨多章节的隐含逻辑冲突识别能力不足，如技术要求与评标标准的隐性不匹配问题。

(三) 优化方向

1) 领域预训练优化：收集新兴行业语料构建专项数据集，开展多阶段预训练，提升专业术语理解能力；

2) 模型轻量化改造：采用 Distil BERT、ALBERT 等轻量化模型，结合模型量化技术，降低推理耗时；

3) 知识图谱融合：构建招投标领域知识图谱，融合语义规则与深度学习模型，提升逻辑冲突识别能力。

五、结论与展望

(一) 研究结论

本文开发的基于深度学习的招投标文件智能编审系统，通过构建领域知识库与优化深度学习模型，实现了合规性检测、完整性校验、一致性比对三大核心功能。实验结果表明：

领域微调后的 BERT 模型在合规性检测中 F1 值达 92.3%，BERT+CRF 模型实体识别 F1 值达 94.1%，Siamese-BERT 模型一致性比对 F1 值达 91.2%，核心算法性能优于传统方法；

系统审查准确率达 92.5%，与人工审查水平接近，单份文件审查耗时降低 80% 以上，显著提升编审效率；

系统能准确定位风险点并提供法规依据与整改建议，为招投标文件编审提供智能化支撑。

(二) 未来展望

未来研究将从三方面推进：

1) 多模态信息处理：融合文本、表格、图片等多模态数据，实现工程量清单、图纸要求等非文本信息的智能审查；

2) 人机交互优化：引入主动学习机制，通过人工反馈迭代优化模型，提升复杂场景适应性；

3) 全流程集成应用：将智能编审系统与招投标交易平台、信用评价系统集成，实现“编审 - 交易 - 监管”全流程数字化闭环管理，进一步推动招投标领域数字化转型。

参考文献

- [1] 中华人民共和国国家发展和改革委员会. 中国招标投标行业发展报告 (2023) [R]. 北京: 中国市场出版社, 2023.
- [2] 王亮, 李娟, 张敏. 基于本体的招投标知识库构建研究 [J]. 情报理论与实践, 2021, 44(8): 145-151.
- [3] 蓝峰杰, 王烈, 黄莹. 一种改进型 Dense-HRNet 和基尼指数动态加权决策的表情识别算法 [J]. 电讯技术, 2022, 62(11): 1683-1690.
- [4] 代磊超, 冯林, 尚兴林, 等. 基于深度网络的快速少样本学习算法 [J]. 模式识别与人工智能, 2021, 34(10): 941-956.