“净语”智能敏感词检测系统策划书

指导老师：李永

参赛人员：姜佳泽 王赛 程关潼 胡玉强 张宇泽

“净语”智能敏感词检测系统

一、执行摘要

在信息爆炸的时代，网络内容安全已成为政府监管、企业合规与品牌声誉的重要方面。传统敏感词过滤工具对谐音、变体、语义伪装等敏感内容检测效率不佳，难以满足日益严格的法规要求。为此，我们推出“净语”智能敏感词检测系统（Jingyǔ Intelligent sensitive word detection system），一款融合高效规则引擎与大语言模型语义理解的智能内容安全检测平台。系统采用 Web + FastAPI + Ollama 的整体设计，使用“AC 自动机+DFA”双规则引擎进行毫秒级初筛，并通过通义千问等大模型对存疑内容进行深度语义分析，实现精准识别、智能判断、结果可复现。“净语”智能敏感词检测系统同时支持实时文本输入与多格式文档（TXT/PDF/DOCX/图片 OCR）上传检测，提供默认模式（高效）与严格模式（深度）双轨并行策略，兼顾性能与精度，广泛适用于平台中多任务、高时效、高敏感场景。

二、产品概述

（一）出发点

1、传统固定敏感检测方法识别“和-谐”、“freedom”等变体表达效果不佳，且无法匹配变体的更新速度；敏感词严格场景下人工复核成本高，反馈速度慢；

2、传统敏感检测方法处理格式单一，多数系统仅支持纯文本，无法解析 PDF、图片等复杂载体；

3、完全基于大语言模型（Large Language Models，LLMS）的检测方法计算成本高，面对大规模请求难以保证实时性，影响用户体验；

4、相关开源项目少，部署难度大。

（二）创新点

1、双重检测机制，使用 AC 自动机+DFA 双引擎匹配机制，支持规则匹配快速筛选 + 存疑内容 LLM 智能检测；

2、两种检测模式，创新使用实时模式和严格模式，实时模式，采用的规则初筛+大模型二次检测，以保证实时要求；严格模式：取消规则匹配快速预筛，所有输入均使用大模型检测，适用于检测率要求高的场景；

3、敏感词库管理，支持敏感词库选择、构建、编辑、移除等功能，以保证敏感词更新变化，同时管理操作简洁友好；

4、该项目支持 TXT、PDF、DOCX、DOC 格式，支持图片 OCR 识别（JPG、PNG、BMP、GIF、TIFF），文件大小限制（10MB），字符限制（10000 个字符），拖拽上传支持，无需依赖第三方工具转换，解决“多格式文档需逐一适配”的效率问题；同时预留格式扩展接口，满足不同场景下的检测需求；

5、创新项目在实验过程中，毫秒级响应时间。单个语句正常响应时间约 5ms，存疑内容单次响应时间约 450ms，连续响应

时间约150ms，适用于实时性要求高场景，优化用户体验和提高工作效率；

6、docker部署+模块化设计+完善的技术文档，便于用户快速部署至不同环境。采用Docker和容器化封装，提供一键部署脚本，无需手动配置依赖，解决跨服务器环境部署不一致问题，确保开发、测试、生产环境一致性，实现快速部署；

7、Web检测界面，设计了简洁美观的 Web 界面，并对主要任务进行分区导航，使操作更加简单；

8、数据安全，检测全程所有数据只在内存中处理，不落盘存储，文件解析后立即释放资源。同时支持支持 HTTPS + JWT 认证，保障传输安全。

三、产品设计

（一）产品定位

“净语”智能敏感词检测系统是一款基于 Web + FastAPI + Ollama 的智能内容安全检测系统，它集高效规则匹配与大语言模型语义理解于一体，面向红山平台和开发者，提供实时文本与多格式文档的敏感信息识别服务，兼顾检测速度、准确率与用户体验。

（二）设计理念

1.核心理念

（1）双重检测: 规则匹配快速筛选 + LLM 智能检测

（2）智能分层处理：默认+严格两种模式，规则初筛+ LLM

精判平衡性能与精度；

（3）高性能: 毫秒级响应时间，支持高并发

（4）可扩展: 容器化部署，支持水平扩展

（5）可维护: 模块化设计，易于维护和升级

（6）高可用: 健康检查，自动重启，故障恢复

（7）全格式覆盖：从纯文本到图片 OCR，打通内容输入边界；

（8）高易用性：拖拽上传、实时反馈、键盘快捷操作。

2.技术原则

（1）容器化: 使用 Docker 实现环境一致性

（2）微服务: 服务分离，独立部署

（3）API 优先: RESTful API 设计

（4）数据持久化: 模型和配置数据持久化存储

（5）监控友好: 完善的日志和监控机制

（三）用户界面原型

敏感词检测系统主要由顶部导航栏和三个核心功能标签页组成：文本检测、文档检测和词库管理。图 3-1 以“文本检测”页面为例。

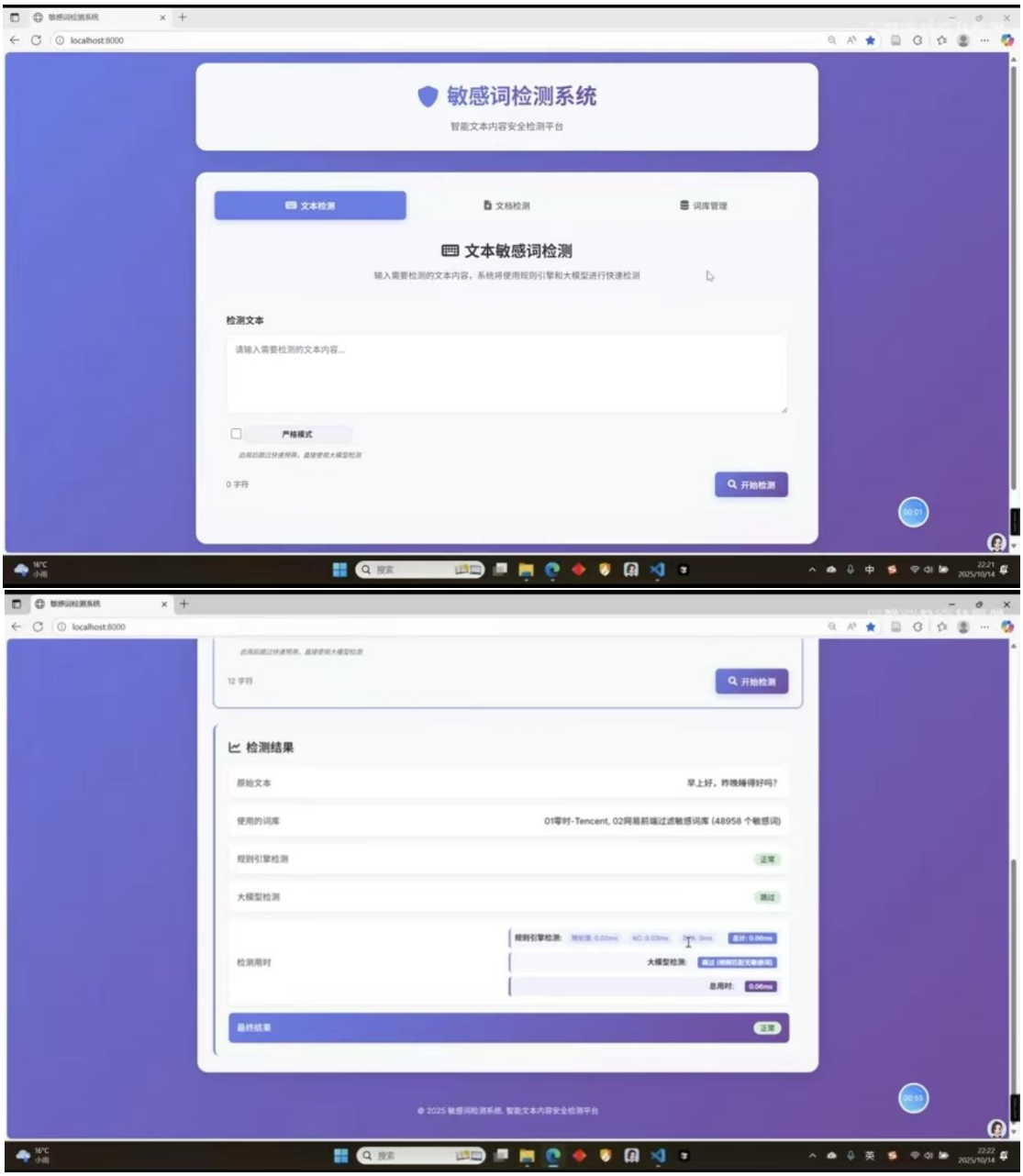


图 3-1 文本检测界面

四、技术方案

（一）整体架构

整体架构如图 4-1：

图 4-1 整体架构图，对“净语”智能敏感词检测系统的核心架构和对应功能进行展示

（二）技术栈选型

技术栈选用及具体技术说明如表 4-1：

类别	技术/工具	说明
后端技术	FastAPI	现代化的 Python Web 框架，支持异步处理和自动生成 API 文档
	Uvicorn	基于 uvloop 的高性能 ASGI 服务器，用于运行 FastAPI 应用
	Pydantic	用于数据验证、序列化和设置管理的库，提升类型安全性和开发效率
	PyPDF2	用于解析和操作 PDF 文档
	python-docx	用于读取和解析 DOCX 格式文档
	antiword	用于解析旧版 DOC 文档的命令行工具
	pytesseract	Python 封装的 OCR 库，用于调用 Tesseract 进行图像文字识别
	Tesseract OCR	开源的图片文字识别引擎，支持多语言识别
	AC 自动机	多模式字符串匹配算法，适用于高效敏感词或关键词检索
	DFA（确定性有限自动机）	用于实现高效的文本过滤与状态机匹配，常用于内容审核或关键词检测
	文本预处理	包括字符归一化、全半角转换、大小写统一、变体统一等，提升文本匹配准确性
前端技术	HTML5	用于构建语义化的网页结构
	CSS3	实现现代化、响应式的页面样式设计
	JavaScript (ES6+)	编写客户端交互逻辑，支持模块化和现代语法
	Fetch API	用于在浏览器中发起异步 HTTP 请求，与后端 API 通信
	Drag & Drop API	实现文件拖拽上传功能，提升用户体验
AI 技术	Ollama	轻量级本地大语言模型运行环境，便于部署和管理 LLM
	Qwen2.5:7b	通义千问 2.5 版本，70 亿参数的量化模型，支持本地高效推理
	Prompt Engineering	通过优化提示词设计，提升大模型输出质量与任务准确性
部署技术	Docker	容器化技术，实现应用及其依赖的隔离与可移植性
	Docker Compose	用于定义和运行多容器 Docker 应用（如后端、前端、数据库等）
	WSL（Windows Subsystem for Linux）	在 Windows 上运行 Linux 环境，便于开发和部署基于 Linux 的应用

表 4-1 技术栈表，对系统各分段使用的技术栈进行介绍和说明

（三）关键算法说明

1.规则匹配引擎的算法架构

在“净语”智能内容安全检测系统中，规则匹配引擎是第一道防线，负责对输入文本进行毫秒级扫描，快速识别明确违规内容。为兼顾匹配效率与规则灵活性，系统采用 AC 自动机 + DFA

双引擎融合架构：

AC 自动机：用于大规模敏感词库的多模式精确匹配

DFA：用于复杂规则（如正则、变体、拼音绕过）的状态机驱动识别。

其总体架构如图 4-2。

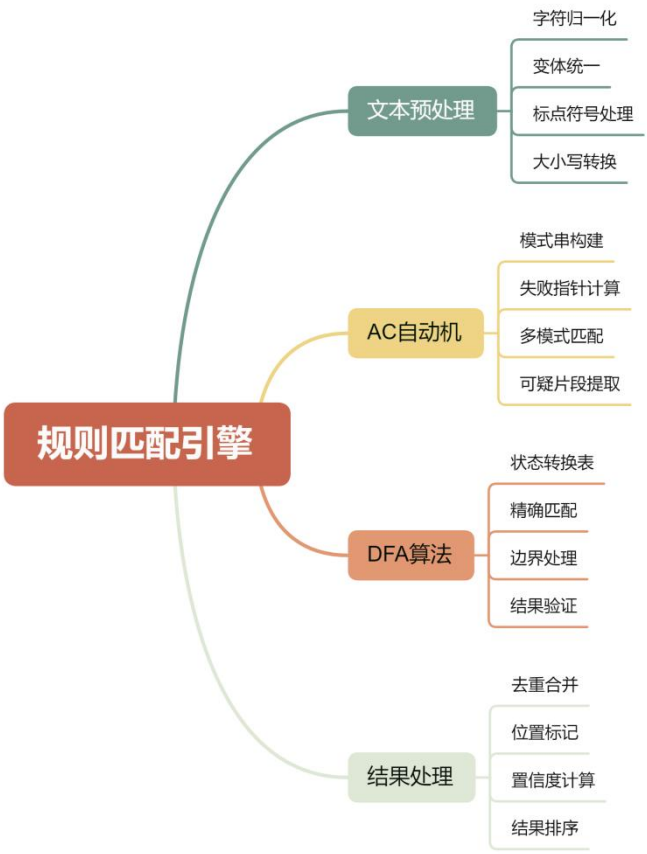


图 4-2 规则匹配引擎的算法架构图

AC 自动机是由 Alfred Aho 与 Margaret Corasick 于 1975 年提出的多模式字符串匹配算法，可在单次文本扫描中同时检测多个关键词，其时间复杂度接近文本长度 O(n)，尤其适用于数万至百万级规模的敏感词库场景。该算法基于 Trie 树结构构建，通过引

入失败指针与输出指针两大机制提升匹配效率，在系统中主要承担支持热更新词库管理、提升检测性能与优化内存占用的作用。

DFA（确定有限状态机）作为一种抽象计算模型，广泛应用于正则表达式引擎与词法分析器。它通过状态转移表驱动，能在固定时间内判定输入是否符合特定模式，尤其适用于处理复杂规则、变体表达及拼音绕过等非标准文本。DFA 由状态集合、输入字符集、转移函数、起始状态和接受状态五个核心要素构成，在系统中主要用于变体识别、正则匹配支持与执行性能优化。

2.AI 服务(Ollama)的算法架构

在“净语”智能内容安全检测系统中，大语言模型（LLM）是规则匹配引擎之后的“智能大脑”，负责对存疑内容进行语义理解、上下文分析与风险意图识别。为实现数据安全、低延迟、可私有化部署的目标，系统推荐采用 Ollama 作为本地大模型运行与管理平台。

其总体架构如图 4-3。

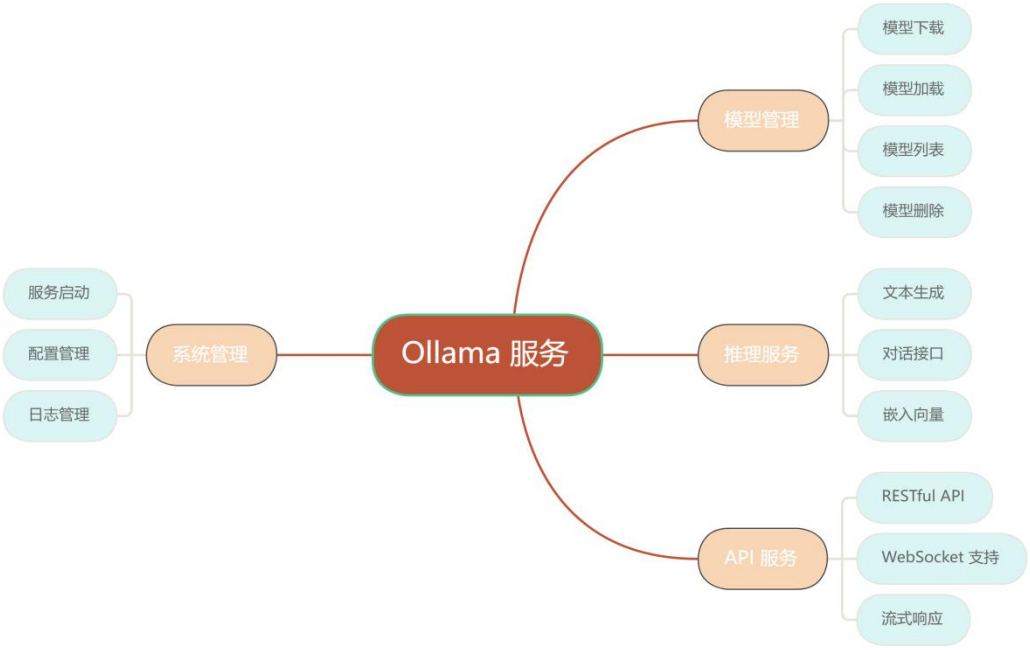


图 4-3　AI 服务（Ollama）的算法架构图

Ollama 是一个开源的本地大模型运行框架，由 Ollama 公司开发，旨在让开发者能够在本地机器或私有服务器上轻松下载、运行和管理大型语言模型，无需依赖公有云 API。Ollama 具有模型即服务，一键拉取主流模型，本地运行、数据不出内网，轻量高效、资源占用可控，RESTful API 接口，支持模型定制与微调等诸多优点。本系统在设计过程中根据实验结果对比，选用 Qwen2.5：7B，使得模型检测速度最大化。

（四）数据流与安全性

1、所有数据在内存中处理，不落盘存储；

2、文件解析后立即释放资源；

3、支持 HTTPS + JWT 认证，保障传输安全；

4、可选开启审计日志（记录操作时间、IP、结果摘要）。

五、功能使用描述

（一）顶部导航栏

顶部导航栏主要对页面进行展示，以及对三大任务进行分区和导航，主要包括两个功能：

1、显示系统的系统标题，标识当前应用；

2、提供标签页切换功能，支持用户在“文本检测”、“文档检测”和“词库管理”之间自由切换。

（二）实时文本检测

该模块用于对用户输入的纯文本内容进行实时敏感词检测，包含五个功能组件：

1.文本输入区域：提供多行文本框，供用户粘贴或输入待检测文本；

2.严格模式选择：提供“严格模式”与“默认模式”切换，控制匹配策略（如是否匹配变体、谐音等）；

3.字符计数显示：实时统计并显示当前输入文本的字符数量；

4.检测按钮：触发敏感词检测流程，启动分析；

5.检测结果展示：以高亮或列表形式展示检测出的敏感词及其位置、所属词库等信息。

进入“文本检测”标签页，在文本框中输入内容（支持粘贴）后，系统自动计数并实时高亮潜在风险词，点击【开始检测】或按 Ctrl+Enter 查看结果，结果共分为六个部分：

一是原始文本，对识别的文本进行显示；

二是使用的词库，显示当前文本检测使用的敏感词库；

三是规则引擎检测，显示使用规则引擎模式检测的结果，并用红绿两种颜色进行高亮处理：

跳过（绿色）：未使用规则引擎检测

正常（绿色）：规则引擎检测无异常

异常（红色）：敏感词个数+具体敏感词

四是大模型检测，显示使用大模型模式检测的结果，并用红绿两种颜色进行高亮处理：

跳过（绿色）：未使用大模型检测

正常（绿色）：大模型检测无异常

敏感（红色）：大模型检测结果异常

五是检测用时，显示规则引擎检测用时、大模型检测用时、总用时，其中规则引擎检测用时包括预处理用时，AC 用时和 DFA 用时；

六是最终结果，文本检测的最终结果，综合规则引擎检测和大模型检测结果，并用红绿两种颜色进行高亮处理：

正常（绿色）：检测无异常

敏感（红色）：检测结果异常

具体功能使用及检测流程如图 5-1：

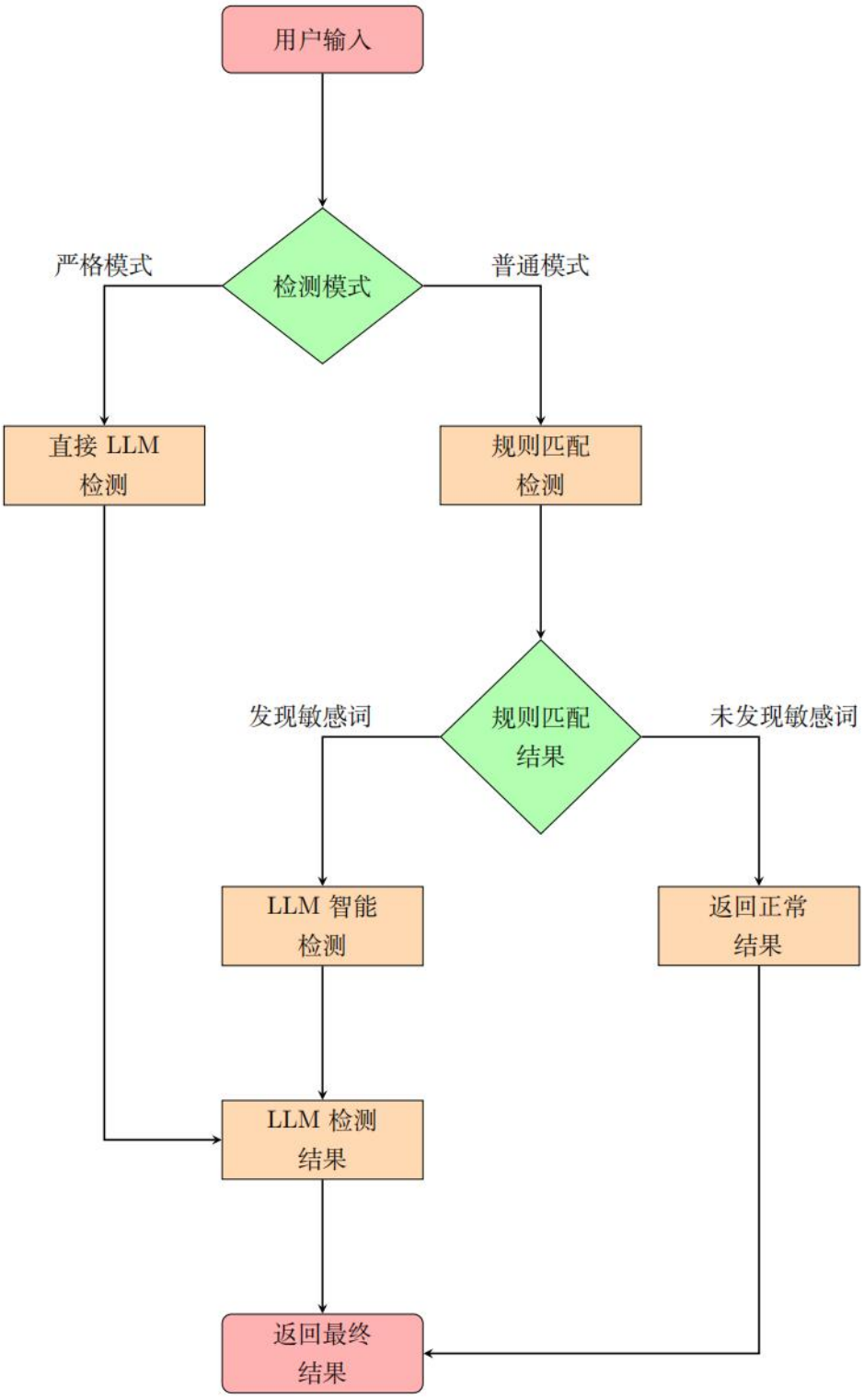


图 5-1 实时文本检测流程图

（三）文档检测

该模块支持上传并检测包含文本的文档文件，具体包括五个功能：

1.文件上传区域：提供文件选择按钮，支持用户选择本地文件；

2.拖拽上传支持：支持将文件直接拖拽至指定区域完成上传；

3.文件信息显示：显示已上传文件的名称、大小、格式等基本信息；

4.检测按钮：启动文档内容解析与敏感词检测；

5.检测结果展示：展示文档中检测出的敏感词列表及所在段落或页码信息。

切换至「文档检测」标签页后，灵活选取上传方式，支持拖拽文件或点击选择文件，同时支持 txt、doc、docx、pdf、jpg、png 等多种格式上传，而后系统自动对文件大小、解析后字符数进行校验：

文件大小 ≤ 10MB

解析后字符 ≤ 10,000

若通过，则开始解析，按照文件类型调用相应的解析工具：

TXT 文档：直接读取

PDF 文档：PyPDF2 解析

DOCX 文档：python-docx 解析

DOC 文档：antiword 解析工具

OCR：pytesseract 文字识别

点击【开始检测】或按 Ctrl+Enter 进入检测流程，默认使用严格模式，检测完成后显示检测报告，共分为八个部分：

一是文件名，显示所检测文档的名字；

二是文件类型，显示为 TXT、PDF、DOCX、DOC 、OCR（JPG、PNG、BMP、GIF、TIFF）；

三是文本长度，对所检测文本进行计数；

四是使用的词库，显示当前文本检测使用的敏感词库；

五是规则引擎检测，显示使用规则引擎模式检测的结果，并用红绿两种颜色进行高亮处理：

跳过（绿色）：未使用规则引擎检测

正常（绿色）：规则引擎检测无异常

异常（红色）：敏感词个数+具体敏感词

六是大模型检测，显示使用大模型模式检测的结果，并用红绿两种颜色进行高亮处理：

跳过（绿色）：未使用大模型检测

正常（绿色）：大模型检测无异常

敏感（红色）：大模型检测结果异常

七是检测用时，显示规则引擎检测用时、大模型检测用时、总用时，其中规则引擎检测用时包括预处理用时，AC 用时和 DFA 用时；

八是最终结果，文本检测的最终结果，综合规则引擎检测和

大模型检测结果，并用红绿两种颜色进行高亮处理：

正常（绿色）：检测无异常

敏感（红色）：检测结果异常

具体功能使用及检测流程如图 5-2。

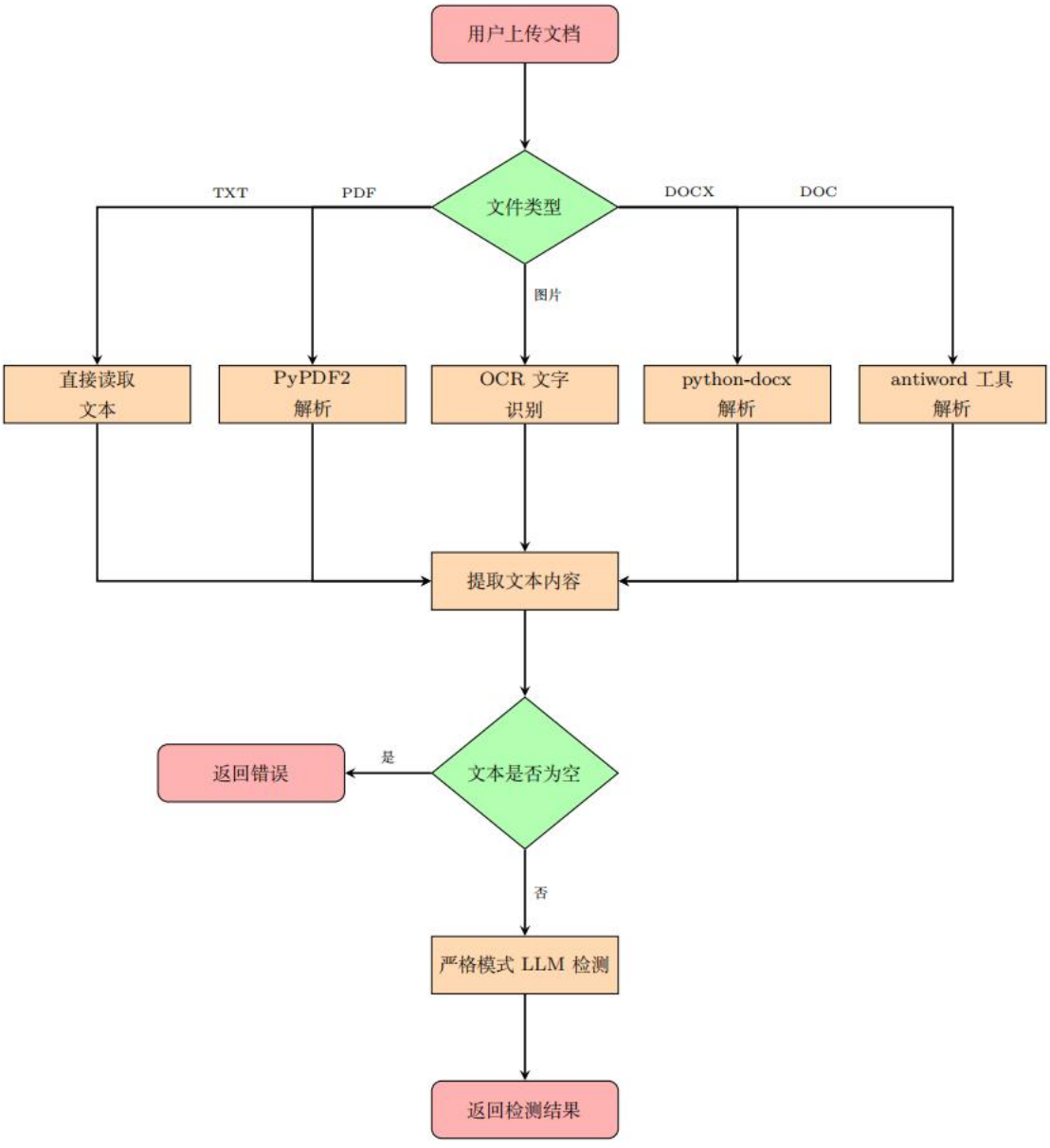


图 5-2 文档检测流程图

（四）敏感词库管理

该模块用于管理和维护敏感词库，分为三个子模块：

1.使用词库列表

已选词库显示：列出当前检测任务中启用的词库；

词库统计信息：显示所选词库的总词数、分类分布等统计信息；

更新检测词库按钮：将当前选择的词库同步至检测引擎，确保检测使用最新词库。

2.词库列表管理

词库列表显示：以列表形式展示所有已创建的词库；

创建新词库按钮：用于新增一个独立的敏感词库；

编辑词库功能：支持对已有词库进行重命名、启用/禁用等操作；

删除词库功能：支持删除不再需要的词库（需二次确认）。

3.词库编辑器

词库名称输入：可编辑当前词库的名称；

敏感词列表编辑：提供文本区域或列表形式，支持批量添加、删除或修改敏感词；

敏感词计数显示：实时显示当前词库中包含的敏感词总数；

保存/取消按钮：保存修改或取消编辑并返回。

六、应用价值

1.精准识别，规则 ＋LLM 双重验证，显著降低误报率与漏

报率；

2.高效处理，AC 自动机实现毫秒级响应，支持大规模内容筛查；

3.格式全面，覆盖文本、办公文档、扫描件、图片，真正“所见即检”；

4.灵活可用，双模式切换适应不同业务场景，兼顾效率与安全；

5.易于集成，提供标准化 API，可嵌入 CMS、IM、审批系统等。

七、核心团队人员介绍

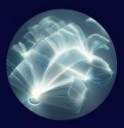
（一）指导老师

（二）参赛人员

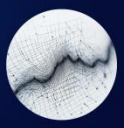
参赛队长

姓名：姜佳泽

基本信息：男 · 2000年生



学历：硕士研究生在读



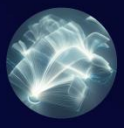
研究方向：深度学习、行为识别

邮箱：2477156418@qq.com

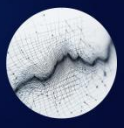
参赛队员

姓名：王赛

基本信息：男 · 1995年生



学历：硕士研究生在读



研究方向：作战数据保障

邮箱： 986133953@qq.com

参赛队员

姓名：程关潼

基本信息：女 · 2000年生

学历：硕士研究生在读

研究方向：作战通信保障

邮箱： 1311244473@qq.com

参赛队员

姓名：胡玉强

基本信息：男 · 2004年生

学历：硕士研究生在读

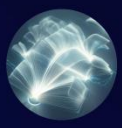
研究方向：深度学习、行为跟踪

邮箱： 3142608087@qq.com

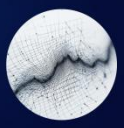
参赛队员

姓名：张宇泽

基本信息：男 · 2005年生



学历：本科在读



本科专业：应用心理学

邮箱：2569679395@qq.com