## 5G敏感信息监测系统

*今天你博学了吗？-黄凯、许力、陈文舟、李忻原、张宁*

### 背景及意义

背景：在传统短信时代，短信仅支持140字节文本格式的内容，对于敏感信息可以通过关键字或者关键字组合等方式识别过滤。而5G消息承载于IP网络，可以支持文本长度大于140的超长文本，同时支持图片、音频、视频、文件、地理位置、卡片消息等更丰富的媒体格式的内容，因此需要更精准可靠的的模型去识别敏感信息。

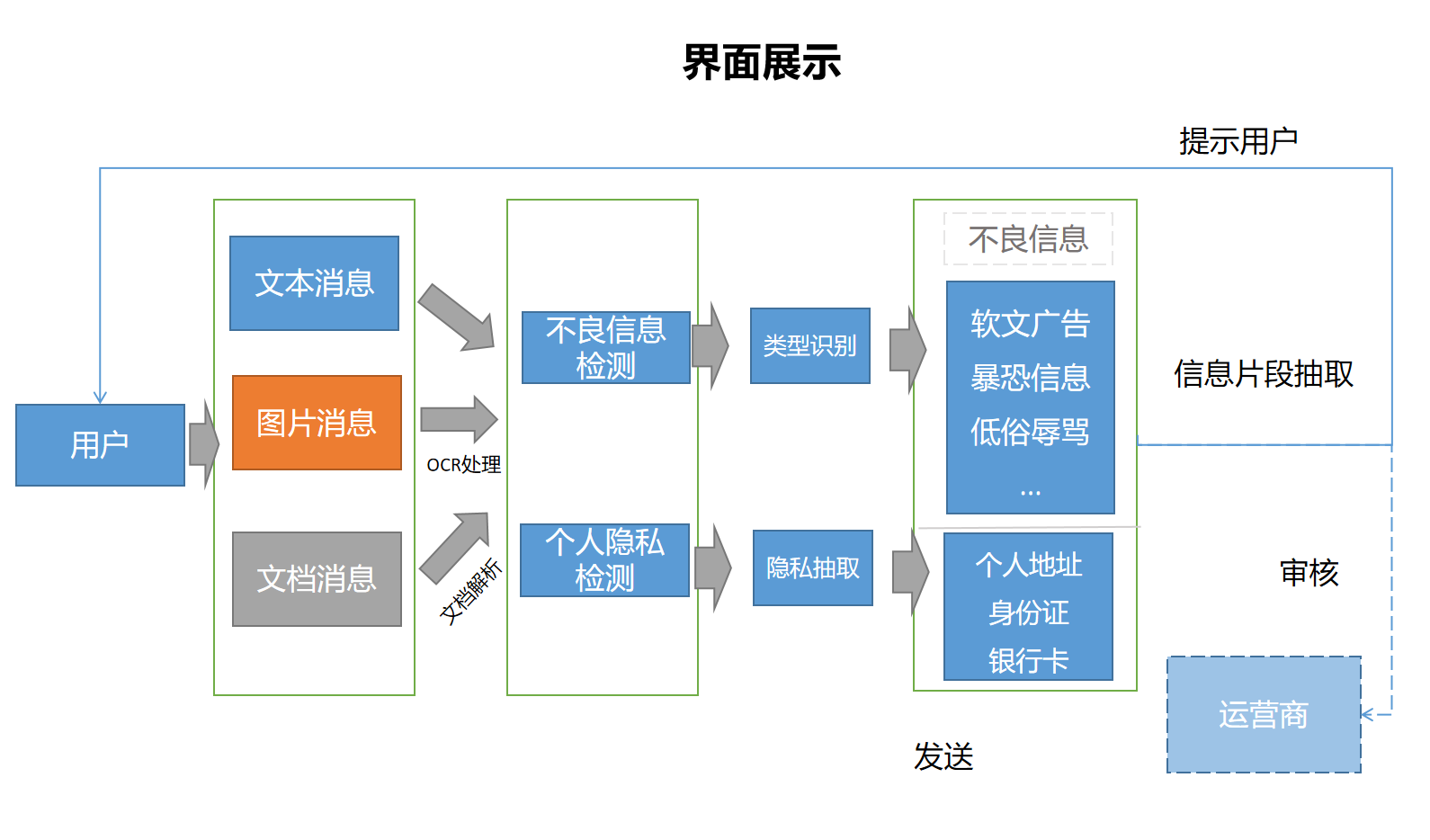
意义：鉴于5G消息的复杂性，对于不良消息内容则需要更加有效的手段来识别和拦截。在消息通信界面中，当识别到敏感信息以后可以进行风险提示；在消息后台管理系统中，敏感信息识别可以辅助人工审核，审核通过后再下发，避免内容和业务提供商的消息因错漏造成大范围的恶劣影响。

<https://zhuanlan.zhihu.com/p/515795622?utm_id=0>

<https://www.zte.com.cn/china/about/magazine/zte-technologies/2021/8-cn/4/5.html>

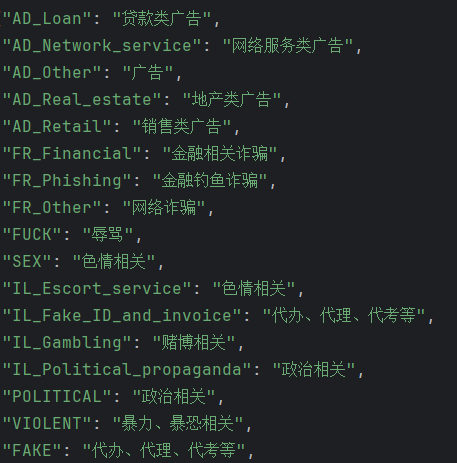
### 系统框架

**描述：**对于超长文本内容，除了通用的关键字扫描识别，还需要增加上下文关联、超长文本分段并发扫描识别等优化扫描和匹配算法，识别更隐蔽的非法内容，加快命中速度。对于多模态信息（**本方案以图片、文件为例**）等多媒体内容，一般采用样本库快速匹配与内容识别相结合的方式。一方面对于已入样本库的文件可以快速识别，另一方面还要保证未入样本库的消息内容的识别。可提供消人工审核，降低不良信息的错漏风险。



系统框架中主要实现了如下几个功能：

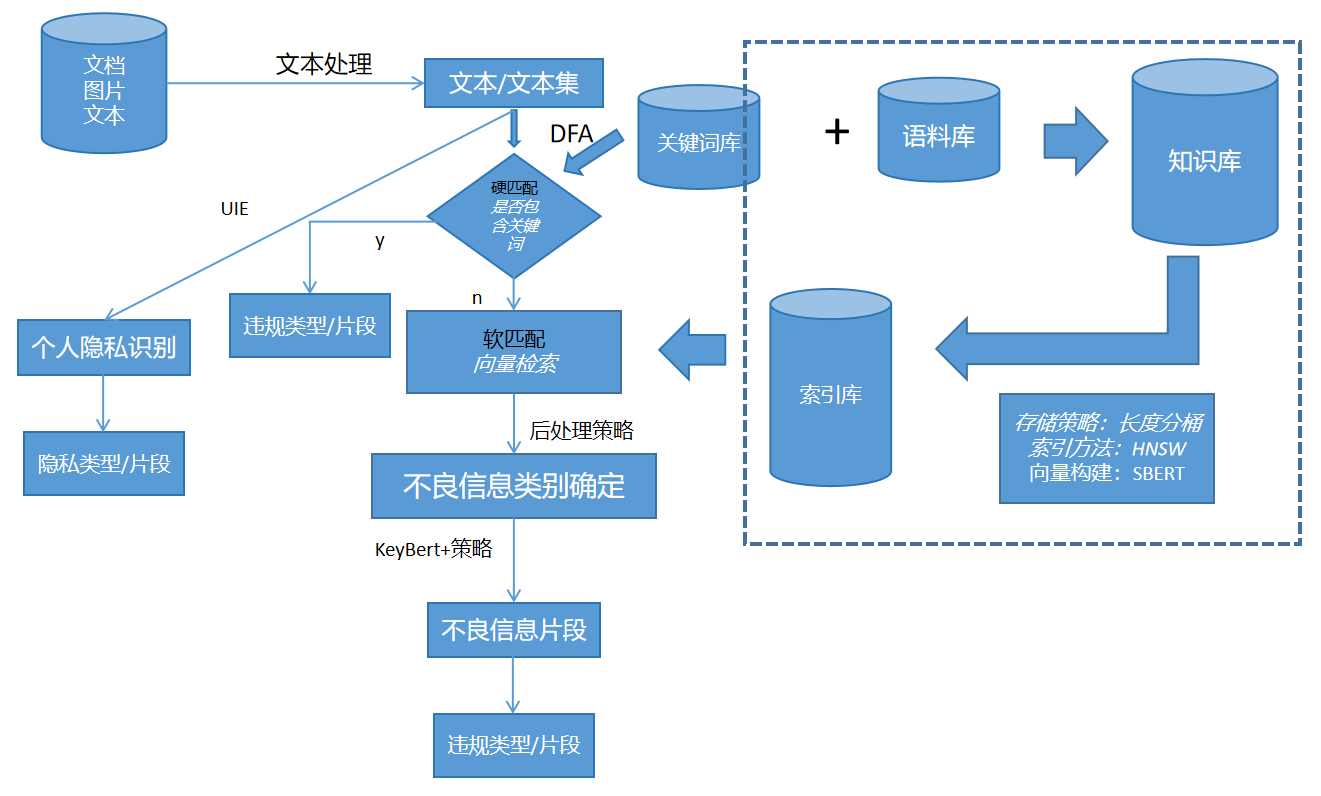
1. 多模态信息的处理,多模态信息包含：
2. 文本；
3. 文档，包括pdf、txt、excel三种格式；
4. 图片，包括jpg、png等格式。
5. 敏感信息识别包括：
6. 不良信息



1. 个人隐私信息：个人地址、身份证号、银行卡号
2. 敏感片段抽取：
3. 不良信息中抽取关键短语，作为界面展示信息；
4. 个人隐私信息抽取，作为界面展示信息
5. 界面展示功能：
   1. 输入文本、上传文档和图片；
   2. 敏感信息分类：分为两个板块展示，分别为不良信息和个人隐私信息；
   3. 敏感片段标亮；
   4. 敏感信息提示。

### 技术实现

基于以上功能，最终采用了如下技术方案，模型整体框架如下：



#### 多模态信息处理

1. 文本处理：包括长文本切割、繁简转换、空格，标点符号过滤等；
2. 图片处理：使用PaddleOCR框架实现文字识别，这里采用PP-OCRv4版本，分别包含检测模型和识别模型；
3. 文档处理：pdf使用PdfReader框架识别，excel文件使用pandas处理。

#### 不良信息识别和片段抽取

##### 不良信息识别

不良信息识别这里采用了**“以搜带分”**的理念，该理念的核心在于将分类任务转化成检索匹配任务，通过维护一个知识库，使用检索算法召回语料，最终可以取top-1或者top-k的众数标签来实现消息的分类。该方案的选用依据如下：

1. 消息的内容具有较强的外延性，正常的消息多种多样，导致覆盖的消息及其有限，

若用多分类算法会导致样本极不均衡，效果欠佳；

1. 使用搜索的方式召回topk个标签，类似于一个knn分类，可以取top1或者众数，理论可行；
2. 可控性较强：在做效果迭代时，分类任务需要收集语料然后重新训练，检索任务则只需维护知识库或者向量索引即可；
3. 语义相似度模型构建相对简单，可以同时结合业务语料和外部语料进行训练。
4. 可快速部署上线，极少量语料既可以完成检索分类任务，在紧急任务和少量语料场景中更高效和友好。

技术方案：

1. 关键词匹配：

采用DFA匹配算法，通过构建字典树，结合每个节点的两种状态（开始状态、结束状态）实现DFA匹配。

1. 检索匹配：
2. 向量检索：
   * 1. HNSW索引结构，hnswlib实现；
     2. 分桶存储，即在划定长度区间存储向量，这样可以提升检索匹配的准确性；
     3. 匹配模型自训练，结合jaccard和语义匹配值构建难负样本，使用SBERT训练匹配模型，预训练模型为ernie3.0然后抽取模型结构，可以实现文本向量化。
3. 检索及后处理策略：
   * 1. 多路检索召回：BM25检索结合向量检索
     2. 后处理：分别取top10类别频次的的众数作为最终文本类别

##### 片段抽取

采用KeyBERT模型实现不良信息的片段抽取，原理如下：复用SBERT提取文档嵌入以获得文档级表示。然后提取n-gram词/短语。最后使用余弦相似度来找到与文档最相似的单词/短语。取topk个最相似的词语并整合作为描述整个文档的词/短语。

#### 3.隐私信息识别和抽取

采用序列标注方式实现地址的抽取ernie3.0-gru-crf

采用正则实现身份证号的抽取和验证

采用正则和luhn算法实现银行卡号的抽取和验证。

*Luhn算法：<https://blog.csdn.net/weixin_33720452/article/details/91862644>*

#### 4.web框架

采用streamlit前后端web框架实现界面展示，这里前后端一体，直接使用内置方法可实现界面模块化，标亮等功能。基础界面如下：

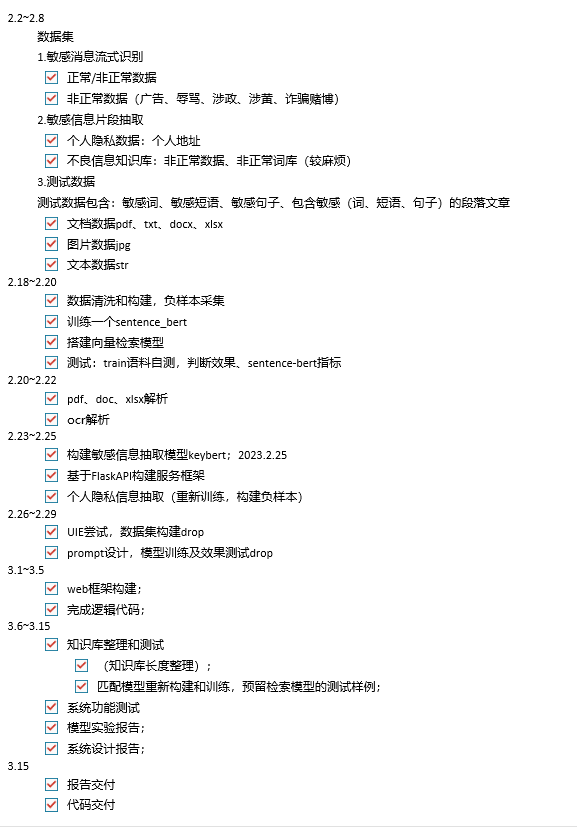


#### 5.API服务搭建

搭建一种基于Flask web框架的API服务框架，将各个任务的功能以API服务的形式提供，可以全局调用功能，较少了脚本的重复调用，提升代码的简洁性和可读性。API功能清单如下：

1. 关键词查找
2. 关键词库的词语添加、更新
3. 文本/知识库向量化
4. 知识库索引更新
5. 文本检索匹配
6. 隐私信息抽取

### 里程碑



### 总结和思考

#### 局限性

1.缺乏业务语料，导致未录入的语义识别效果不高；

2.缺乏有业务经验的组员，会导致对业务理解不够充分，功能和实际运用可能有偏差

3.课题时间比较紧凑，一个月的时间规划，节奏非常快，很多先进的模型未能验证。

#### 2.和大模型对比

1.资源占用小：6B大模型约占12G显存，若没有足够的资源则大模型适用性不高。

2.较高的实时性：本课题以实际落地运用为主，充分考虑了模型的响应效率，因此在工程实践上设计了API接口及模型复用等功能。而大模型响应时间较长，目前还不具备实时/近实时的响应能力。

#### 3.长期规划

本课题初步完成敏感信息检测demo，但是仍然有如下几个方面值得去探究：

技术方面：

1. 会话级的的敏感信息识别
2. 识别模型的精细化（包含表情、变体字、敏感图像等内容的处理和识别）
3. 语音、视频等更多模态的识别

业务方面：

1. 敏感信息识别差异化：比如接收的信息和发送的信息关注的敏感类型不同
2. 敏感类型多样化：结合业务场景，可能会有其他类型的敏感信息
3. 个人隐私细化：目前只检测地址、身份证号、银行卡号，依据隐私数据等级不 同可以对个人隐私的定义和类型进行调整。

类似应用场景：

1. 舆情分析
2. 客服质检和服务质量评估