## 5G消息敏感信息监测系统

### 背景及意义

背景：传统短信仅支持140字节文本格式的内容，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。在传统短信时代，对于垃圾短信可以通过关键字或者关键字组合等方式识别过滤。

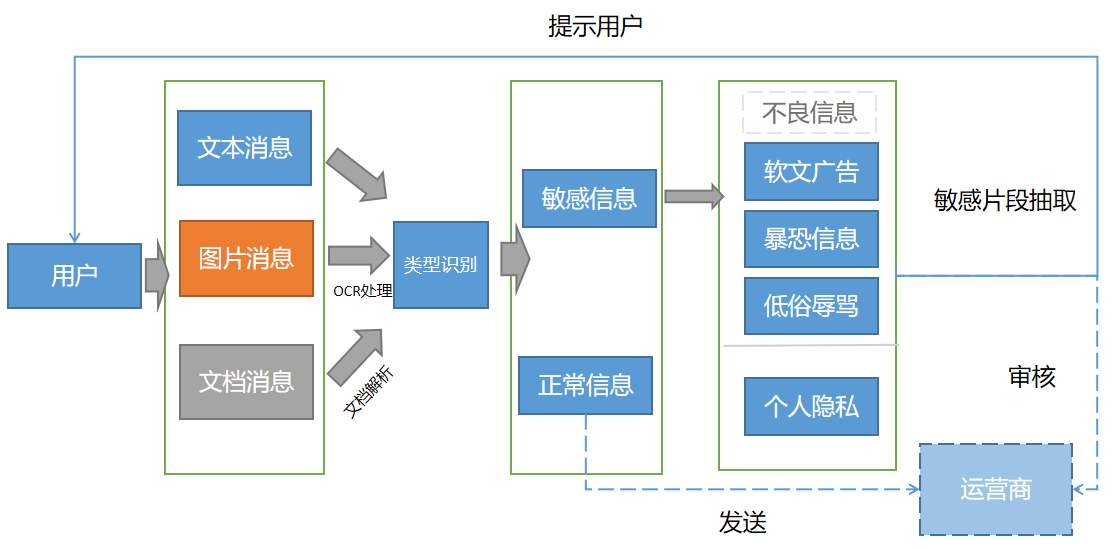
意义：建议5G消息的复杂性，对于不良消息内容则需要更加有效的手段来识别和拦截。此外，CSP（内容和业务提供商）的消息影响范围更广，所以对于CSP的消息内容需要采用更严格的审核机制，如消息内容须人工审核通过后再下发，避免因错漏造成大范围的恶劣影响。

<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>

### 模型框架

**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，同时增加人工审核，降低不良信息的错漏风险。



### 功能细节描述

1. 图片消息：OCR；
2. 文档解析pdf、txt、docx、excel：
   1. <https://blog.csdn.net/assember/article/details/133142593>；
   2. <https://blog.csdn.net/liaozp88/article/details/129850787>
3. 敏感信息包含：隐私信息和不良信息；
4. 前后端框架选择：streamlit
5. 文本消息识别：流式分类：正常/非正常-->非正常的类别（广告、暴恐、辱骂、反动）
6. 敏感片段抽取方法（提升可解释性）：
   1. 隐私信息：命名实体识别+正则策略，然后标亮；
   2. 不良信息：分隔符分割，数据库向量检索返回最大值或者超过一定阈值然后标亮（非监督）；

### 技术方案

技术框架： Paddle+Meilisearch+Streamlit

1. 语料收集和整理

流式语料集及标签的整理：

1.正常的消息

2.非正常消息

* 1. 不良信息（广告、辱骂、涉政、涉黄）
  2. 个人隐私信息（住址、身份证号、银行卡号）

片段抽取知识库语料：

不良信息+敏感词

1. 信息处理
2. ocr实践

参考框架：paddleocr

其他方案：<https://blog.csdn.net/qq_27900321/article/details/124445542>

<https://blog.csdn.net/liaozp88/article/details/129850787>

1. 文档解析:txt、doc、excel解析
2. pdf解析：pdf-->图片-->文字（复用ocr框架）
3. 消息流式识别
   1. 消息是否包含敏感信息：二分类 bert 系列 --> ernie2.0/3.0 +nn
   2. 敏感信息的类型：多分类 ernie2.0/3.0 +nn

要求：准确率>95%\*95%=90.25%

1. 敏感片段抽取（麻烦，需要不断调试）
   1. 个人隐私：NER+正则 bert+bilstm+crf +正则 or UIE

<https://github.com/skyfaker/chinese_address_extract>

<https://github.com/qq1065507891/ChineseAddressNER/tree/main>

* 1. 不良信息：关键词检索-->消息切分检索（切分策略）（较麻烦）

1. Web选型及任务（doing）

选型：Python streamlit框架 [官网链接](https://docs.streamlit.io/)

任务：

1. 敏感信息标亮
2. 返回文本提示信息：该条信息疑似广告/xx信息/包含个人敏感信息，请谨慎操 作
3. 功能测试及报告
4. 与大模型框架对比
5. 资源占用小
6. 较高实时性

考虑可落地性，在保证准确性的前提下，使用了轻量级的推理模型和检索框架。与 大模型框架相比响应时间短，具有较高的实时性。

### 排期



### 长期规划

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

可应用场景：

舆情分析

客服质检和服务质量评估