智能安审堡垒在消息类业务中的应用

邵佳梦，马昭征
（中国电信上海号百信息服务分公司，上海 200050）

摘　要：在当今时代，技术架构的不断演进与自然语言处理技术的深入发展，共同推动了消息类业务的全面升级。便捷的新型平台架构和精准高效的文本识别算法，不仅极大地提升了面向用户服务的智能化水平，也显著提高了风险识别技术识别风险的准确度。立足于消息类业务，特别是短消息类业务在防欺诈、防骚扰、防不良信息等领域的实际应用，旨在对传统方法进行创新。提出了一种新型的文本安全审核架构，并针对当前普遍存在的异形字问题，创新性地融合多种算法，提出了全新的解决思路。通过构建智能安审堡垒平台，实现了消息类业务端口的精细化管理与异形文本的合规审核，采用微服务架构设计，既利于平台能力横向扩展，也利于服务能力纵向叠加，从而为消息类业务的稳健发展提供了坚实的支撑。

关键词：短消息；安审；算法；防欺诈
中图分类号：TP393
文献标志码：A
doi: 10.11959/j.issn.1000−0801.2025079

0　引言

随着互联网行业的迅猛发展，每天有数十亿条文本消息在网络上流转，市场规模庞大。然而，这一现象也伴随着色情、赌博、毒品以及政治敏感内容在网络上的泛滥，使得互联网企业面临着巨大风险，对文本安全审核的需求日益迫切。以行业短消息客户为例，在多租户模式下，运营商往往难以控制租户的行为。租户发布的涉黄、涉诈、涉敏感信息等问题，可能导致主端口账号被封停，影响整个业务的运营。尽管面临着这些风险，许多企业在账号被封之前并未采取有效措施降低业务风险。因此，不仅运营商需要安全审核，所有从事文本信息转发业务的企业对此都有巨大需求。业务提供商（service provider，SP）消息类业务涉及多个行业，具有高实时性和广泛的用户覆盖面，这些特点使得对内容合规性和发送审核的要求更为严格。

此外，经过调研，现在违法违规短消息的隐蔽程度越来越高，其中不乏出现通过拆分字、形似/音似字、表情符号等来传递违法信息的案例，使得传统的拦截系统无法识别。根据工信部发布的相关报告[1]中的数据，2023年上半年全国移动短消息业务量为9 346.5亿条，其中共拦截各类垃圾短消息超过90亿条，约占短消息业务量的1%。其中涉诈短消息占比约为16%，利用变体字生僻字实施短消息诈骗的占比约为涉诈短消息的64%，即10.2%左右，占总体短消息业务量的0.1%，即9亿条。因此，结合前沿技术来提升短消息文本的审核效果不仅运营商的迫切需求，也是所有开展文本信息转发业务企业的需求，市场潜力同样非常大。

本文提出了一种全新的消息类智能安审方法，即在消息云网关（message cloud gateway，

收稿日期：2024−11−23；修回日期：2025−02−25

MCG）上建立安审堡垒系统应用，以实现更灵活、效果更好的消息类文本安全审核服务。该应用聚焦于文本消息的安审重点，比如文本内容安审、异形字识别、端口报备情况、签名报备情况、黑白名单等。该应用具有两个显著特点：一是采用功能独立的微服务模块构建，服务与服务之间可以叠加，用户能够选择自己需要的能力，实现个性化、灵活的消息类业务安全审核；二是结合变体字算法集，通过对算法的设计与组合，提升其对异形字的识别能力，弥补传统文本识别无法识别拆分字、音似字、形似字的缺点，提升隐蔽违法消息的筛查效果。

1 变体字算法介绍

针对文本识别问题，传统的文本识别方法适应性差、需要分离训练目标，预分割和后处理阶段操作较为麻烦。在计算机行业飞速发展的今天，自动处理算法逐渐成熟，文本检测和识别算法的准确度都大大提升[2]。在本应用中，为了实现复杂异形字的识别，采用了4种相关算法。

1.1 拆分字变体算法

在检测不良信息的过程中，有时存在拆分字干扰检测结果的情况。例如，将阵字拆分成“阝”“车”。如果一个句子中含有过多的拆分字，会导致原有的模型无法识别该条语句的语义，因此需要首先进行预处理，对于输入的信息，将其中的拆分汉字识别出来后，转换为拆分之前的正常汉字，之后再把处理过的语句送入后续的检测模型。

该预处理功能能够识别语句中的拆分字，并将拆分字还原成原来的正常汉字。本算法采用知识库匹配的方法来完成该任务。首先整理了拆分字数据集，包含1 686个汉字以及它们拆分后的形式。将该数据集读入一个字典存储。在查找并判断拆分字时，遍历整个输入内容，查找是否有与字典中拆分后的元素相匹配的字段，若有则替换为对应位置的汉字。

1.2 形似字变体算法

给定短消息文本，本算法能够对于其中的形似字进行识别与还原。遵循分阶段的纠错架构，其主要包含以下步骤。

（1）错误检测模块

- 先进行混淆形近词典匹配，例如，对“高梁——高粱”等词进行识别与纠错。
- 进行常用字典匹配，认定切词后不在常用词典中的词为疑似错词。
- 使用*n*元语法（*n*gram）语言模型，即通过判定某个字的前后搭配2gram和3gram的似然概率值是否低于句子文本平均PPL（perplexity，衡量语言模型性能的一种指标，反映了模型对语言序列的预测能力）值来判定是否为错词。

（2）候选召回

- 使用混淆集词典进行错字替换。
- 使用形似字词典进行错字替换。

（3）候选排序

基于统计语言模型KenLM（一个高效的语言模型库）计算句子似然概率，取概率值超过原句且最大的句子。

1.3 音似字变体算法

对于给定的短消息文本，对于其中的音似字进行识别与还原。音似字变体通常指的是与原汉字发音相似但含义却不同的汉字变体，与形似字同为干扰正常文本阅读与机器文本分类的手段，其识别与纠错往往与形似字采用同样的措施与手段。音似字算法与其形似字实现思路基本相同，区别在于数据词典的选择。

1.4 干扰符号变体算法

在不良信息检测这个任务中，存在干扰信息混淆检测内容的情况，导致训练好的检测模型无法识别不良信息。这种情况一般是预训练数据集中这类情况出现较少所导致的。所谓的干扰信息

可以分为如下几类。第一类是 emoji 表情，一种现在流行在网络上的表情符号，可以清晰地表达发信人想传达的意思，但传统检测方法可能无法判断。第二类是用符号代替的汉字，如用 V 来代替“微”。第三类是将电话号码或者网页统一资源定位系统（uniform resource locator，URL）拆分后夹在文本中，避开模型的检测。

本算法包含预处理模块，能够将输入语句中的干扰信息剔除，或者单独提出干扰信息作检测。对于 emoji 的检测，直接调用上述库中的 emoji.emojize 方法，可以将语句中所有 emoji 符号转换为中文含义。对于颜文字与特殊符号的检测，收集到了一个有 1 600 个颜文字对应其含义的对照表，在颜文字检测中，可以首先提取语句中的特殊符号再对照查表，判断是否为表中的颜文字，若是则替换为中文含义。对于 URL 与电话号码的识别与提取，采用正则匹配的方式，先设定电话号码与 URL 的匹配格式，再对输入进行处理。对电话号码，可以先提取出输入中所有数字部分，再与之前设定好的格式进行匹配，判断是否为一个合法号码。对 URL 采用与上述相同的思路，首先提取非中文的字段，然后进行格式匹配。

2　应用架构设计与功能

安审堡垒整体架构如图 1 所示。

业务模块分为账号管理模块、企业账号管理

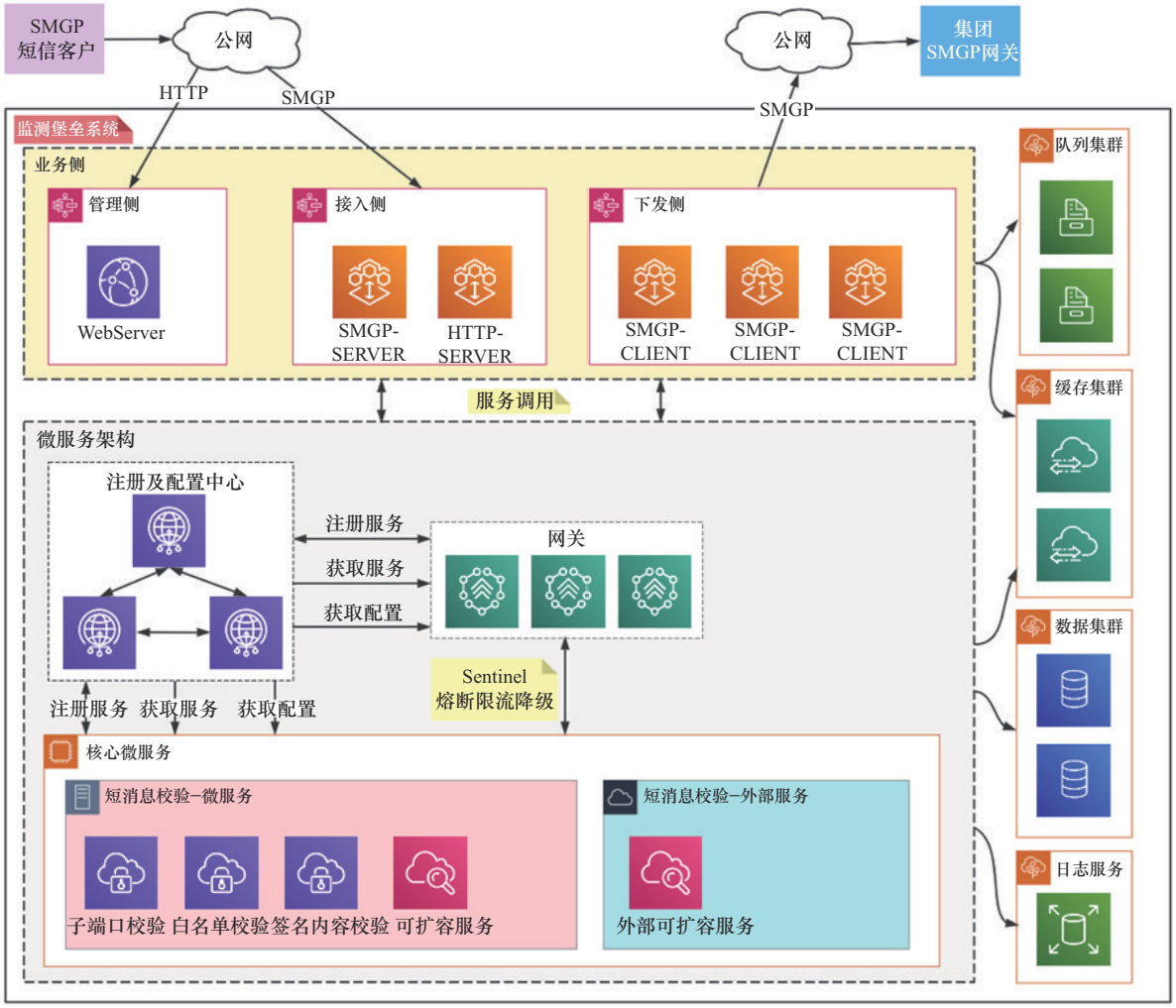

图 1　安审堡垒整体架构

模块、企业信息管理模块、安审模板管理模块、扩展码号、签名管理模块、审核管理模块、安审模块、日志管理模块、统计模块。

后期规划提供应用程序接口（application program interface，API），支持调用外部能力实现安审子模块功能，或将本系统安审子模块开放给其他系统调用。

系统架构主要分为四大模块，接下来进行详细介绍。

2.1 管理及签约模块

2.1.1 账号管理模块

用于管理平台登录账号，账号类型分为3种：系统管理员、管理员、操作员。

- 系统管理员：具备角色管理、企业账号管理、企业账号管理、企业信息管理、码号管理、签名管理、审核管理、日志管理、统计查询等功能权限。
- 管理员：具备企业账号查询、企业信息查询、码号管理、签名管理、审核管理、日志管理、统计等权限。
- 操作员：具备日志管理、统计查询等权限。

2.1.2 企业账号管理模块

通过公网Web（网络）界面开放给行业短消息用户，用于进行管理企业账号、修改密码等操作。

企业账号权限包括对本企业主端口号扩展码号进行号段规划配置，以及对该号段使用的签名进行增加、删除、修改、查询。

2.1.3 企业信息管理模块

与集团备份企业信息同步，包括企业全称、主端口号、企业白名单因特网协议（Internet protocol，IP）地址，一旦绑定就不可修改。所有企业初始状态默认为新增规则需审核，为满足高信用企业的操作时效性，可由系统管理员配置为“自动审核”，并保留任务日志备查。系统管理员可在此模块新增企业账户、同步企业信息、选择安审模板。

2.1.4 企业账号安审能力签约模块

企业账号安审能力签约模块，由管理员根据业务需求，对该企业需要进行的安审模块，即原子能力进行签约和配置。企业用户通过公网Web对签约的安审模块进行参数配置，并适用于业务。

2.1.5 安审模板管理模块

灵活配置企业所属安审子模块，提供子接口安审功能选项和签名校验功能选项。可扩充其他安审手段，如黑名单功能选项、短消息校验功能选项、5G消息富媒体安审功能选项、语义分析等。模块提供API，支持调用外部能力实现安审子模块功能，或将本系统安审子模块开放给其他系统调用。

2.1.6 审核管理模块

新增规则将生成一条待审核任务，由系统管理员和管理员审核生效。系统管理员可通过修改企业信息，对信用度高的用户配置“自动审核”机制。

2.2 短消息存储转发模块

2.2.1 SMGP-SERVER

采用短消息网关协议（short message gateway protocol，SMGP）接口与用户进行连接，接收短消息，解析短消息，获取短消息内容及其他信息，并缓存短消息业务流，存入队列。

2.2.2 SMGP-CLIENT

按照下游能力配置每秒获取的最大消息数，并进行处理，完成削峰。根据企业用户签约的安审模块和配置参数，进行校验。校验成功则通过SMGP接口转发集团，获取集团接口响应后，将响应的流消息存入处理提交响应的队列（SubmitResp），在此期间不会生产消息，而是全部透传。校验失败，则封装SMGP消息存入SubmitResp队列。接收回执消息，并将其存入处理回执或确认

消息的队列（Receipt）队列。

2.2.3　队列模块

队列模块采用卡夫卡集群方式进行。

2.2.4　分布式缓存模块

缓存模块采用基于内存的高速键值存储数据库远程字典服务（remote dictionary server，Redis）集群方式进行，完成长消息的缓存校验和消息内容的解耦。

2.2.5　数据库模块

该模块能够持久化企业用户信息、安审能力签约信息和系统运行的配置信息等。

2.2.6　日志及统计模块

该模块能进行发送量统计，支持以月、日、小时、分钟为参数区间，统计发送总量，最小颗粒度为分钟，有总发送量和各企业发送量两种统计方式。此外，该模块能够统计平均发送量，支持统计月平均、日平均、小时平均，最小颗粒度为小时，有总发送量平均和各企业发送量平均两种统计方式。

此外，该模块还能够动态显示每秒建立呼叫数量（call attempts per second，CAPS），分为所有企业合并的CAPS和每个企业单独呈现的CAPS；能够统计各企业发送成功率、拦截率、错误码等。

该模块采用分布式搜索和分析引擎（elasticsearch，ES）集群方式，收集话单消息和日志消息，采用Kibana（数据可视化和分析平台）进行统计的展现。

2.3　微服务管理模块

2.3.1　动态服务发现、配置管理和服务管理平台（Nacos）注册中心

用于发现、配置和管理微服务。Nacos提供了一组简单易用的特性集，帮助您快速实现动态服务发现、服务配置、服务元数据及流量管理。

2.3.2　微服务调度

根据用户短消息安审的签约，选取短消息需要进行的微服务组件进行调用。组件调用可并行进行。

2.3.3　微服务网关

提供了以下几点能力。

（1）提供了统一访问入口，降低了服务受攻击面积。

（2）提供了统一跨域解决方案。

（3）提供了统一日志记录操作，可以进行统一监控。

（4）提供了统一权限认证支持。

（5）提供了微服务限流功能，可以保护微服务，防止雪崩效应发生。

2.4　微服务能力模块

2.4.1　企业IP地址校验服务

对企业绑定的IP地址、企业身份标识号码（identity document，ID）、主端口号进行校验。IP不属于该企业用户，或与主端口不匹配，则返回失败错误码。

2.4.2　扩展号码校验服务

通过用户配置的子端口段进行号码校验。子端口不属于企业配置号码段，则返回失败错误码。

2.4.3　签名管理校验服务

通过配置子端口的短消息内容的签名，对短消息内容进行校验，符合校验内容的则透传，否则返回失败错误码。

2.4.4　内容审核服务

采用传统敏感词拦截和异形字算法拦截相结合的模式，针对文本检测是否涉及黄赌毒等违规内容开放公网点对点接入，可对敏感信息直接拦截，并返回错误码。支持管理员Web页面自定义敏感词，对包含自定义敏感词的信息进行拦截，并返回错误码。

3　应用输出内容

本文结合不良文本算法集，提供文本内容智能安审，输出分析类型，并按需输出报告。内容

安审输出结果包含如下所示的内容。

变体干扰短消息还原后内容，例如zun傲白勺硌戶→尊敬的客户。

可疑URL提取，如33哈7，75啦j.c就n→33775j.cn。

可疑号码，如2197q3q8p1078→219XXXX078（qq号码），158m361v7v3v474→158XXXXX474（电话号码）。

短消息类别及其概率如下。

- 涉黄概率，20%。
- 涉诈概率，50%。
- 涉政概率，70%。
- 涉毒概率，60%。
- 涉赌概率，80%。
- 谩骂概率，60%。

可疑交易信息，包括银行卡号、开卡行、卡类别、疑似收款人、金额，例如，疑似银行卡ID——9988XXXXXXXX8071；开户行信息——XX银行（借记卡）；开卡类型——借记卡；开卡长度——16；疑似收款人信息——刘向阳；疑似组织公司信息——平安银行；疑似转账金额信息——500元。

4 微服务模块的优势

本文所述应用可以在安审堡垒中建设具备安审能力的签约适配器，根据客户的不同信用等级，签约不同的能力审核机制；建设了可扩充的短消息监管能力池，应对未来未知的短消息潜在威胁，如图2所示。此处举例进行说明。

（1）信用等级良好的A用户，可以仅签约签名校验能力，安审堡垒对A用户发送的信息只验证签名，验证通过后就把正确的信息提交给集团网关。

（2）信用等级一般的B用户，可签约签名校验和内容校验能力，安审堡垒对B用户发送的信息会既进行签名校验，也进行内容校验，全部通过才会提交到集团网关。

（3）当出现了新的安全问题，安审堡垒支持扩容，在能力池中新增新的文本类审核能力，并提供签约，即可动态地支持新能力的应用。

在此基础上，继续开发新的堡垒功能，提供文本类进阶安审服务，供短消息业务体系平台调用，适用于一切文本类素材的安审，包括短消

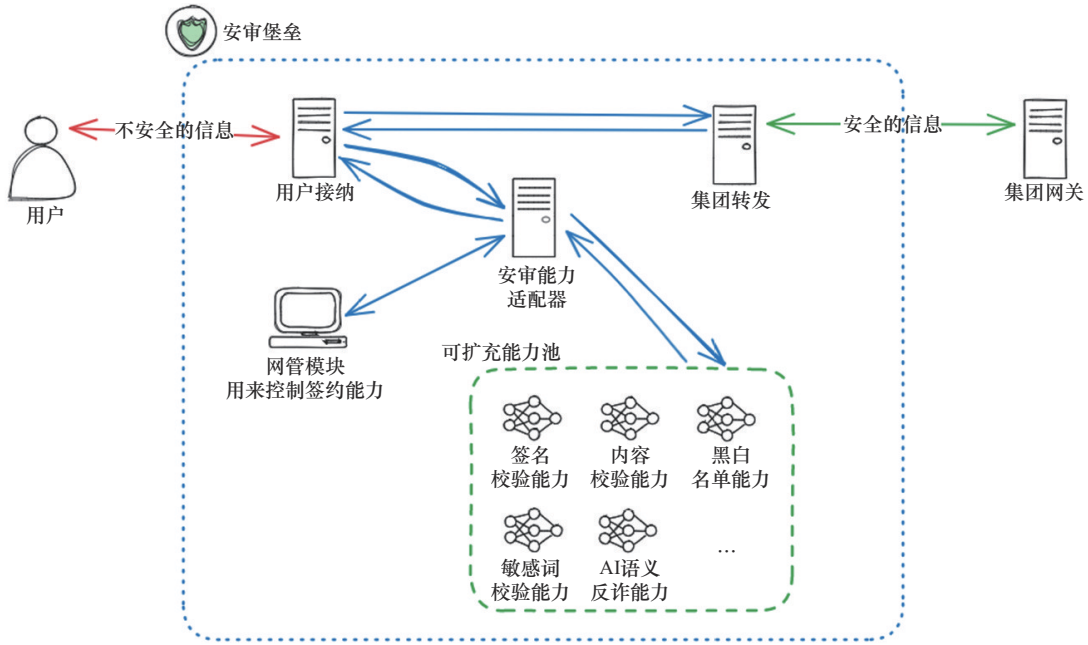

图2 安审堡垒对风险短消息的解决方案

息、文本乃至智能质检中语音识别技术（automatic speech recognition，ASR）输出的文字，甚至于直接输入docx，PDF等格式的文档，输出定制化的安审结果。提供多种微服务化安审能力，可以以签约模式针对自身各种业务定制微服务套餐，包括内容审核、端口签名校验、敏感词稽核、组合敏感词稽核、短链安审、拆字安审、形似字、音似字安审、风险评级等等。

5　结束语

本文深入探讨了技术架构的演进与自然语言处理技术如何共同推动消息类业务的全面升级，并提出了一种新型的文本安全审核架构。通过对短消息类业务在防欺诈、防骚扰、防不良信息等领域的实际应用进行分析，我们创新性地解决了异形字问题，并构建了智能安审堡垒平台。这一平台不仅实现了消息类业务端口的精细化管理，还确保了异形文本的合规审核，采用了微服务架构设计，为业务的横向扩展和纵向叠加提供了可能。

综上所述，本文的研究成果为消息类业务的稳健发展提供了坚实的理论和技术支撑。我们相信，随着技术的不断进步，文本安全审核的系统架构将进一步完善，为运营商消息类业务的健康发展保驾护航。在此，我们期待业界同仁的关注与参与，共同推动自然语言处理技术在消息类业务中的应用，为构建更加安全、智能、高效的消息服务生态贡献力量。最后，希望本文的研究成果能够为相关领域的研究和实践提供有益的参考，为我国信息技术产业的发展做出应有的贡献。

参考文献：

[1]　工业和信息化部. 2023年通信业统计公报[EB]. 2024.

[2]　宫法明, 刘芳华, 李厥瑾, 等. 基于深度学习的场景文本检测与识别[J]. 计算机系统应用, 2021, 30(8): 179-185.

[作者简介]

邵佳梦（1987– ），男，中国电信上海号百信息服务分公司研发工程师，主要研究方向为短消息相关能力建设和应用。

马昭征（1990– ），男，中国电信上海号百信息服务分公司产数开发工程师，主要研究方向为人工智能技术与应用、系统集成与开发。