基于AI的消息业务内容安全治理解决方案及关键技术

Message Service Content Security Governance Solutions and Key Technologies Based on AI

赵晨斌[1]，陈浩然[2]，李善诗[2]，李蔚然[2]（1. 中国联合网络通信集团有限公司，北京 100033；2. 中讯邮电咨询设计院有限公司，北京 100048）

Zhao Chenbin[1], Chen Haoran[2], Li Shanshi[2], Li Weiran[2]（1. China United Network Communications Group Co., Ltd., Beijing 100033, China; 2. China Information Technology Designing & Consulting Institute Co., Ltd., Beijing 100048, China）

摘　要：

调研了消息业务内容安全治理现状以及基于AI的消息业务内容安全治理的关键技术，并根据当前技术短板，提出了优化建议。根据消息业务网络架构特征，对短信和5G消息2种消息业务提出了不良消息治理的方案。通过AI辅助垃圾短信策略生成、异常电子信息提取、AI稽核拦截明细，赋能短信业务治理，通过构建AI能力基座，利用深度学习赋能5G消息治理。

关键词：

短消息业务；富媒体消息业务；5G消息；内容安全治理

doi：10.12045/j.issn.1007-3043.2024.08.002

文章编号：1007-3043（2024）08-0008-05

中图分类号：TP181

文献标识码：A

开放科学（资源服务）标识码（OSID）：

Abstract：

It investigates the status quo of content security governance of message business and the key technologies of AI-based content security governance of message business, and puts forward optimization suggestions according to the shortcomings of current technologies. It proposes bad message governance schemes for SMS and 5G message services according to the architecture characteristics of message services. Through AI-assisted generation of spam SMS policies, extraction of abnormal electronic information, and AI auditing and interception details, SMS business governance is enabled. By building AI capability base, 5G message governance is enabled by deep learning.

Keywords：

Short message service; Rich content service; 5G messaging; Content security governance

引用格式：赵晨斌，陈浩然，李善诗，等. 基于AI的消息业务内容安全治理解决方案及关键技术[J]. 邮电设计技术，2024(8)：8-12.

0 前言

随着移动通信服务的快速发展，运营商消息类业务在为用户提供便捷通信服务的同时，也为不良消息的传播提供了渠道。根据12321受理中心通报，2023年第4季度垃圾短信的投诉共计4.3万件，环比上升17.0%，同比上升3.9%。同时，5G消息等新型富媒体消息通信业态的出现，为消息安全治理带来新挑战。如何应用AI技术赋能消息内容安全治理，解决传统的黑白名单机制、关键词策略、富媒体消息哈希值对比等治理手段存在的短板，成为亟待研究的问题。

收稿日期：2024-06-14

1 消息业务安全现状及痛点

1.1 消息业务发展现状

电信运营商消息业务主要包括短信、彩信、数字短信以及5G消息。消息内容治理主要运用包括消息内容分析、黑名单机制、多媒体文件的相似度分析等技术，在短信中心、彩信中心、5G消息中心侧实现对不良消息的监控与处置。随着大数据、人工智能等新技术的不断发展，将AI等新技术运用于不良消息治理成为新趋势，这能够有效提升拦截监控的及时性、精准性，并降低人工稽核成本。

1.2 需求痛点

1.2.1 文本消息

对于文本消息治理，在文本内容对抗过程中，传统的技术手段是利用敏感词库、文本特征库、违规样本库等风险规则库，并通过简单的比对来判定短信是否违规。但文本变形（如异体字、花体字、火星文）技术门槛低，不良消息容易通过文本变形的技术手段规避平台监管。此外，通过脚本可以迅速将垃圾短信恶意文本传播曝光，并不断进行文本进化，这对攻防响应的实时性要求较高。如果采用关键词组合及流量阈值的方式进行内容识别封堵，极易造成正常用户被误拦截，导致短信无法正常发送，进而引发用户投诉甚至造成舆情[1]。

1.2.2 富媒体消息

5G消息支持多种富媒体消息形态，可以在不加好友、免注册、免关注的情况下向被叫用户发送消息。随着5G消息的普及，富媒体消息的传播量会大量增加，传播速度更快，用户行为模式更复杂，不良富媒体消息量也随之增加；同时，内容生产技术的发展使得富媒体内容的演化加剧，形式、种类更加丰富多样。传统的富媒体内容治理主要是将富媒体内容与违规样本库中的样本进行比对，若与样例一致或高度匹配，则判定消息为违规消息，这种治理方式泛化能力较弱、效率低，并且要维护海量的样本数据库，需要高昂的成本，很难适应富媒体消息业务的发展需要。

2 不良消息内容治理关键技术

2.1 文本消息不良内容治理关键技术

文本消息不良内容治理主要有3个环节：文本预处理、文本特征提取及文本相似度比对。文本预处理主要针对垃圾短信的各种变体进行数据清洗及文本归一化操作，包括特殊编码替换、同音形音替换、形近字替换及其他文本（拼音、表情符号）归一化操作；文本特征提取通过语义理解算法对归一化的文本内容去除停用词、提取切分词、输出文本哈希特征、输出垃圾短信分类。

文本相似度对比中常用的距离计算包括余弦相似度、欧式距离、曼哈顿距离、切比雪夫距离等，也可以将文本向量看成不同的多维变量，使用统计相关系数进行相似度计算，如皮尔逊相关系数和斯皮尔曼相关系数等。文本相似度算法主要有传统的基于字符串、基于统计和基于知识库的语义文本相似计算。基于字符串的计算方法包括最小编辑距离方法、最长公共子序列方法、N-Gram模型方法以及Jaccard系数与Dice系数；基于统计的计算方法是将文本转换成一个向量，计算表征文本向量间的距离，包括基于向量空间模型、潜在语义分析主题模型、隐含狄利克雷分布模型；基于知识库的计算方法是运用结构化语义词典或互联网知识，发掘概念信息和概念间层次关系，或利用网页内容间链接关系进行相似度计算。目前主流的实现方式是基于AI深度学习的语义文本相似度计算，可通过孪生网络模型架构以及基础上的交互模型来进行相似度度量[2]。

2.2 富媒体消息不良内容治理关键技术

富媒体消息不良内容治理有2个环节：内容预处理和富媒体内容治理。

2.2.1 内容预处理

内容预处理的目的在于将富媒体内容转化为更便于系统分类和识别的文本、图片和音频，包括转换格式、改变文件大小、降噪处理、分解、分类、规范化等。例如，在处理视频文件时，首先会将文件转换为特定格式，再分解并提取出字幕、视频图像和音频内容。音频内容的处理则需要使用长短期记忆网络（LTSM）对音频做降噪处理[3]，然后按照旋律、呼吸、娇喘、语言等属性对音频进行分类，其中语言类音频则通过自动语音识别技术（ASR）转化为文字内容。

2.2.2 富媒体内容治理

预处理后，对于文本内容，调用文本内容治理接口进行处理，富媒体内容则被传送至富媒体内容治理模块进行治理。

现有的AI辅助多媒体内容治理主要采用图片匹配策略，主要使用的方法有3种。一是直方图法[4]，该方法通过统计图像中不同像素点的个数，得出像素分布的直方图，使用相关性计算方法计算2张图片像素点分布直方图之间的相似性以进行图片相似度对比。二是向量法[5]，该方法将图片转换为向量形式，通过计算两向量的余弦相似度来判断2张图片是否相似。三是哈希法[6]，该方法将整张图像或图像中的局部特征点转换为一个固定长度的二进制编码，即哈希值，通过计算对比2张图片或2张图片特征点的哈希值距离，判断2张图片是否相似。传统的图片处理方法比较成熟，具有较高的精度，资源消耗较低，但是泛化能力较低，需要维护大量的样本库，效率较低。通过AI深度学习可以弥补现有手段的缺陷。深度学习使用

通过海量数据进行预训练的模型，将特征提取和规则学习统一起来，自动从原始数据中提取出有价值的深度特征并主动学习，通过计算深度特征之间的相似度实现富媒体内容的深度对比，便于处理更加复杂、抽象的问题，减少了传统机器训练中人工提取特征的繁琐过程[7]。深度学习的特征提取能力更强，泛化能力更强，可以处理复杂的问题，但其内部处理机制比较复杂，并且需要海量的训练数据才能达到较好的效果[8]。

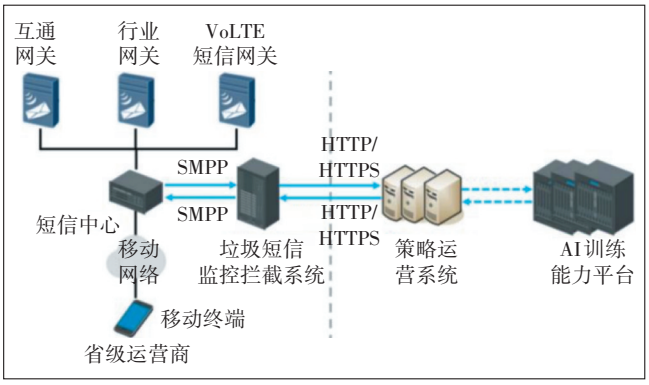

图1　垃圾短信治理组网

3 不良消息治理方案

目前电信运营商短彩信和5G消息业务的网络架构差异较大，根据网络架构的特征，需要制定差异化的短信和5G消息治理方案。

3.1 垃圾短信治理网络方案

AI赋能的垃圾短信治理整体系统包括短信中心、垃圾短信监控拦截系统、策略运营系统、AI训练能力平台4类网元，其组网如图1所示。

其中，短信中心实现所有短消息的存储转发功能。垃圾短信监控拦截系统根据主叫黑白名单策略、发送频次策略、关键词(字)策略等结果，向短信中心反馈短消息鉴权指令。策略运营系统对接垃圾短信监控拦截系统，提供推理服务，对可疑样本数据进行判别并反馈结果，为垃圾短信监控拦截系统提供处置依据，同时可针对样本进行关键电子要素抽取、策略生产及优化建议、数据稽核支撑等操作，对于黑白名单策略、发送频次策略、关键词(字)策略等传统治理策略无法准确识别判断的样本，也可基于语义理解能力进行判断，并确定最终处置方式。AI训练能力平台与策略运营系统对接，负责数据管理、微调训练、模型评估、模型部署等，利用人工打标后的数据，持续训练优化模型以不断提升其准确率。

AI模型辅助垃圾短信治理主要体现在3个场景：AI辅助垃圾短信策略生成、异常电子信息提取以及AI稽核拦截明细。

AI辅助垃圾短信策略生成是指用户发送的短信经过短信采集模块并以策略的方式(关键词+阈值)进行过滤，命中策略的垃圾短信通过AI算法分析后，生成垃圾短信治理策略，具体流程如图2所示。

异常电子信息提取是指对于命中采集策略的垃圾短信，经过预处理与AI模块，提取垃圾短信中包含

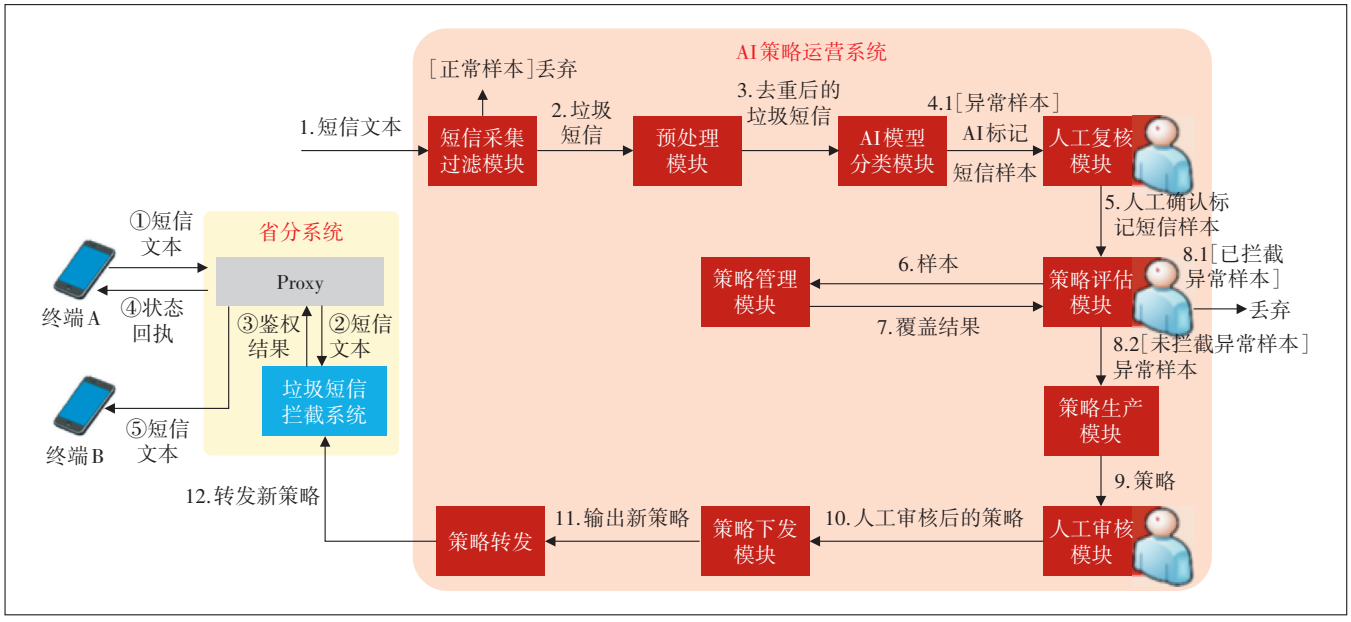

图2　AI辅助垃圾短信策略生成流程

的电子信息并推送至黑网址/域名/IP系统中对其进行鉴别，将被鉴别为异常的电子信息地址反馈给总部垃圾短信策略集中管控平台并生成对应的拦截策略（见图3）。

AI稽核是指将每日拦截短信明细经去重后，通过AI模型模块进行分类，以人工稽核的方式对误判结果进行确认并生成统计报表（见图4）。可通过统计报表找出误拦策略并给予优化意见，实现对垃圾短信低质量策略的优化。

3.2 5G不良消息治理方案

5G消息治理包括对行业消息的治理及个人消息的治理，5G消息治理系统基于AI能力层，打造两级架构的服务平台治理体系（见图5）。AI能力层为服务层提供基础AI能力，包括自然语言处理、相似算法计算、深度学习、智能调优等能力。服务层一级平台负责5G消息全局治理的管理和控制，可用于治理策略实施和运营的全过程，根据二级平台反馈的数据和信息、维护和运营策略，完成治理效果评估，并将治理方案同步给二级平台。二级平台分大区部署，与各大区5GMC及MaaP[9]对接，负责各个大区的行业消息

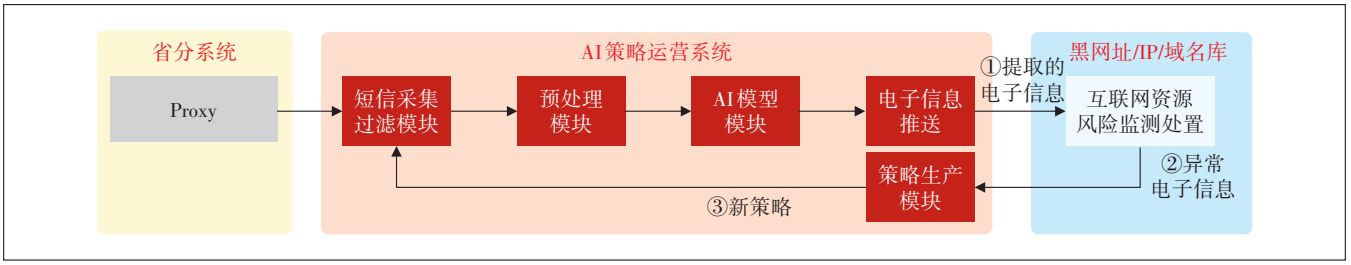

图3　AI异常电子信息提取流程

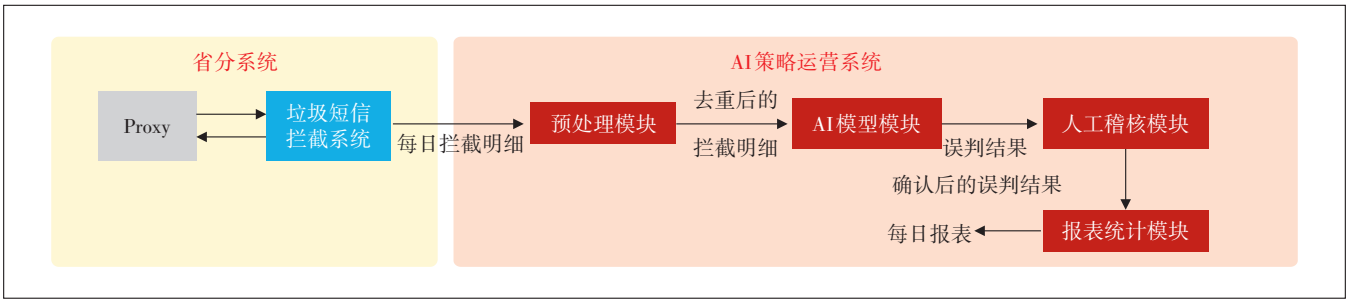

图4　AI稽核流程

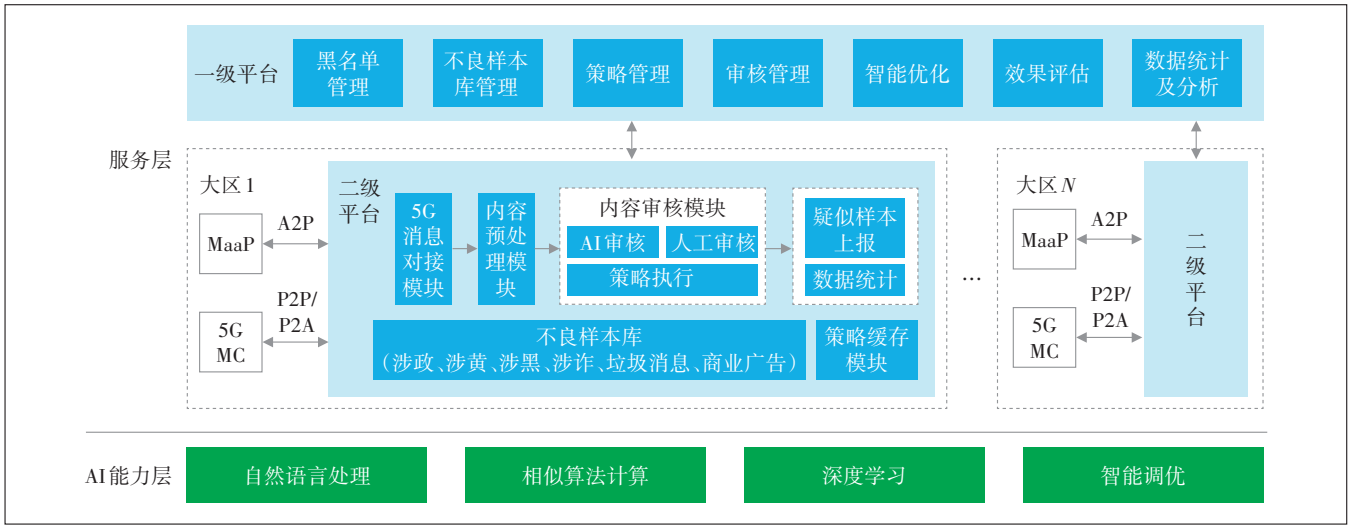

图5　5G消息治理系统架构

（A2P）和个人消息（P2P、P2A）的审核过滤。

AI赋能5G消息治理的关键流程如下。

a）智能优化。对于现网治理过程中发现的各类治理需求，通过AI策略模型智能分析策略的实际价值，对无效策略、重复策略和策略盲点进行调优。

b）内容预处理。二级平台获取消息内容后，根据内容治理需求将需要审核的内容通过AI辅助完成标准化处理，例如将视频文件转化为字幕文本、图片和音频的文件组合，并将处理后的内容同步给内容审核模块进行内容审核。

c）内容审核过滤。内容审核模块采用AI审核和人工审核，二级平台将审核结果返回给MaaP及5GMC，用于判断是否拦截消息。

4 应用效果评估

4.1 AI赋能的垃圾短信治理效果

AI策略运营系统已在某省级运营公司上线，相比传统人工策略运营，AI策略运营在策略生产数量、提质增效、成本节省上效果显著。上线后垃圾短信治理团队仅需3人；日均处理短信号码量从583个提升到1 353个，日均提取策略量从12提升到40，月均误拦量从3 450降到270，月均投诉量从280次降到23次。

4.2 AI赋能的5G消息治理效果

目前，5G消息治理在实验室环境中进行了充分论证，通过深度神经网络，在多媒体内容识别上，该方法能够处理更加复杂的问题。经实验室对比验证，5G消息治理的误判率显著降低，升级前总数据共10 000条，误判975条，误判率为9.75%，升级后总数据共10 000条，误判80条，误判率降低至0.8%。多模态大模型给富媒体内容治理带来了巨大的推动作用，但现有违法违规对象识别分类样本的不足（目前类别仅有1 000类[10]）以及真实场景的复杂性，仍是最大的挑战。

5 总结和展望

传统不良消息治理手段在AI的助力下能完成更高复杂度的语义、语音、图片、视频的内容识别与判断，进而降低误拦漏拦概率，实现治理水平升级[11-13]。本文列举了AI赋能消息治理的关键技术并针对短信治理和5G消息治理提出了差异化的改造适配方案，其中短信治理方案依托垃圾短信监控拦截系统的智能化升级，面向全国31省输出高质量策略；5G消息治理则是基于AI能力基座，具备总部-大区两级平台消息治理能力。解决消息业务在发展过程中产生的业务安全治理新难题，探索如何更好地运用AI技术实现精准治理，需要从业人员在运营过程中持续积累、不断精进、巩固完善，及时跟踪业内相关技术的发展动态，与自身治理现状相结合，实现治理能力的升级[14-15]。

参考文献：

[1] 张凯，张旭. 大数据安全治理与防范[M]. 北京：人民邮电出版社，2023.
[2] 韩程程，李磊，刘婷婷，等. 语义文本相似度计算方法[J]. 华东师范大学学报（自然科学版），2020(5)：95-112.
[3] 李梦晨. 基于人工智能的互联网内容审核模式研究[D]. 北京：北京邮电大学，2022.
[4] 芦俊，从卫华. 一种基于直方图相似度的快速图像匹配算法[C]// 浙江省信号处理学会2013学术年会论文集——信号处理在海洋. 杭州：浙江省信号处理学会，2013：222-225.
[5] 刘冰，李文书. 基于余弦相似度的指纹匹配算法的室内定位方法[J]. 科技通报，2017，33(3)：198-202.
[6] LOWE D G. Distinctive image features from scale-invariant keypoints[J]. International Journal of Computer Vision，2004，60(2)：91-110.
[7] 王华溢，黄要诚，蔡波. 基于传统方法与深度学习方法的图片相似度算法比较[J]. 计算机系统应用，2024，33(2)：253-264.
[8] 杜刚，张晨，杜雪涛. 5G消息服务中的内容安全风险与应对技术[C]//5G网络创新研讨会（2020）论文集. 北京：TD产业联盟，2020：87-90.
[9] 中国通信标准化协会. 5G消息 总体技术要求：YD/T 3989-2021[S/OL]. [2024-05-16]. https://hbba. sacinfo. org. cn/stdDetail/5464c027db0a16062c4823c16f0f00ee886539981d89da560efc7d03cec75272.
[10] 王惠茹，李秀红，李哲，等. 多模态预训练模型综述[J]. 计算机应用，2023，43(4)：991-1004.
[11] 车万翔，刘挺，秦兵，等. 基于改进编辑距离的中文相似句子检索[J]. 高技术通讯，2004，14(7)：15-19.
[12] 王振振，何明，杜永萍. 基于LDA主题模型的文本相似度计算[J]. 计算机科学，2013，40(12)：229-232.
[13] 江敏，肖诗斌，王弘蔚，等. 一种改进的基于《知网》的词语语义相似度计算[J]. 中文信息学报，2008，22(5)：84-89.
[14] 郑志蕴，阮春阳，李伦，等. 本体语义相似度自适应综合加权算法研究[J]. 计算机科学，2016，43(10)：242-247.
[15] 石磊，王毅，成颖，等. 自然语言处理中的注意力机制研究综述[J]. 数据分析与知识发现，2020，4(5)：1-14.

作者简介：
赵晨斌，工程师，学士，主要从事电信反诈、商用密码、信息安全领域政策研究与技术研究工作；陈浩然，高级工程师，硕士，主要从事电信反诈、核心网音视频技术的研究工作；李善诗，工程师，硕士，主要从事5G消息业务及解决方案应用的研究工作；李蔚然，工程师，学士，主要从事电信反诈、短消息业务治理技术的研究工作。