

DOI: 10.13955/j.yzyj.2021.06.07.06

基于人工智能技术的邮政舆情监测分析研究

胡晓菁, 曲春歌, 冯媛

[中邮信息科技(北京)有限公司, 北京 100031]

摘 要: 基于社交媒体平台舆情传播的现状, 运用互联网技术获取公众对于快递企业服务水平的舆情信息, 利用深度学习的自然语言处理算法, 建立舆情关注点分类等模型, 分析了大众对于快递企业服务的正负面舆情和关注的重点问题种类等, 为促进邮政寄递业务服务品质的提升提供数据支撑。

关键词: 舆情; 社交媒体平台; 深度学习; 文本分类

中图分类号: F61 **文献标识码:** A

网络舆情是社会公众对现实生活中的突发热点、焦点事件通过新闻评论、BBS 论坛、微博、贴吧、新闻跟帖及转帖等方式发表自己的言论和态度。随着网络技术的不断发展和创新, 网络舆情已经渗透到了社会生活的各个方面。网络舆情传播对企业的品牌和发展具有重要的意义和作用。如何通过大数据技术手段对企业舆情进行分析管控, 成为邮政部门管理者和研究者亟待解决的问题。

1 研究意义和目标

1.1 研究意义

本文关注邮政舆情监测分析研究, 主要基于以下三个方面的考虑: 一是通过对网络舆情的监测分析, 有助于提升邮政企业品牌形象。通过对舆情信息全方位、多维度的监测分析, 可以促进邮政企业全面及时地监测及处理网络投诉、维护良好的企业舆论环境。二是通过对网络舆情的监测分析, 有助于减少人工监控处理舆情的成本。通过大数据技术, 可实现自动监测重点社交媒体平台公众舆论走

向, 并自动判断舆论情感倾向, 与人工监测舆情相比, 舆情监测在准确性、全面性和时效性上可实现质的飞跃。三是通过人工智能技术建立的舆情监测分析体系, 有助于企业减少负面舆情爆发的风险, 对洞悉舆情态度、舆情管控, 有效预防及减轻企业爆发舆情危机等具有重要意义。四是通过网络舆情监测分析, 有助于提升客户对邮政服务的满意度。通过从客户舆论的视角审视邮政服务问题、短板和差距, 有针对性地制定改进措施, 促进邮政服务品质的持续提升。

1.2 研究目标

本文采用互联网数据获取技术、基于深度学习的自然语言处理技术对社交媒体平台上的邮政及其他快递企业的舆情信息进行获取与全方位的洞察分析, 定位邮政在同行业中所处的舆情地位以及邮政服务的短板。通过互联网技术, 获取公开媒体平台上公众对快递企业的舆论数据, 为舆情分析打下数据基础; 通过基于深度学习的自然语言处理技术, 自主研发基于迁移学习的深度学习方法, 建立

获奖情况: 2020 年全国邮政企业科技创新成果奖二等奖。

作者简介: 胡晓菁(1972~), 女, 江苏常州人, 硕士, 经济师, 主要从事邮政大数据等新技术跟踪和应用创新研究; 曲春歌(1990~), 女, 黑龙江伊春人, 硕士, 工程师, 主要从事邮政大数据等新技术跟踪和应用创新研究; 冯媛(1990~), 女, 山西运城人, 硕士, 经济师, 主要从事邮政大数据等新技术跟踪和应用创新研究。

收稿日期: 2021-05-08

本刊网址: yzyj.sjzpc.edu.cn

正负面舆情分类模型、舆情关注点分类模型,同时建立舆情传播分析模型、舆情对标分析模型、重要舆情转发特征分析模型、客户投诉分析模型等,形成一套完整的人工智能社交媒体平台舆情监测分析体系,帮助企业在舆情环境中占据主导地位,维护企业良好形象,提升品牌传播效率。

2 相关人工智能算法

2.1 数据采集

大数据采集指从传感器和智能设备、企业在线系统、企业离线系统、社交网络和互联网平台等获取数据的过程。数据包括 RFID 数据、传感器数据、用户行为数据、社交网络交互数据及移动互联网数据等各种类型的结构化、半结构化及非结构化的海量数据。大数据采集要保证数据采集的可靠性和高效性,同时避免重复数据。

网络数据采集指通过网络爬虫或网站公开 API 等方式从网站上获取数据信息。该方法可将非结构化数据从网页中抽取出来,以结构化方式将其存储为统一的本地数据文件,支持图片、音频、视频等文件或附件的采集,附件与正文可以自动关联。在大数据时代,网络爬虫是从互联网上采集数据的有利工具。目前已经知道的各种网络爬虫工具有上百个,基本可以分为3类:一是分布式网络爬虫工具,如 Nutch;二是 Java 网络爬虫工具,如 Crawler4j、WebMagic、WebCollector;三是非 Java 网络爬虫工具,如 Scrapy (基于 Python 语言开发)。

2.2 自然语言处理(NLP)技术

NLP 是研究人与计算机交互的语言问题的一门学科,是语言信息处理的一个分支,也是人工智能的核心课题之一。

NLP 经历了基于规则方法、基于统计方法、深度学习三个阶段,通常包括四大类常见任务:分类任务(文本分类/情感分析等)、序列标注(分词/词性标注/NER/语义标注等)、句子关系判断(QA/自然语言推理等)、生成式任务(机器翻译/文本摘要等)。

本研究涉及的 NLP 任务为文本分类任务,即根据给定文档的内容或主题,自动分配预先定义类别标签,从而实现输入公众舆论文本数据后,自动判定其所属的情感类别标签。

2.2.1 基于机器学习的文本分类方法

由于汉语语言的复杂性,传统方法对中文文本的情感分类存在如下四个难点。

一是分词不准确:中文句子由单个汉字组成,通常第一个要解决的问题就是“分词”。但由于汉字组合的歧义性,分词的准确率一直难以达到完美,而不准确的分词结果会直接影响最终分析。

二是缺乏标准完整的情感词库:与中文相比,英文目前有相对完整的情感词库,对每个词语标注了比较全面的情感类型、情感强度等,但中文目前缺乏这样的情感词库。

三是否定词问题:例如“我不是很喜欢这个商品”和“我很喜欢这个商品”,核心情感词都是“喜欢”,但是两个句子却表达了相反的情感。这种否定词表达的组合同样非常丰富,即使将分词和情感词库的问题彻底解决,针对否定词否定范围的分析也是一个难点。

四是不同场景和领域的难题:部分中性的非情感词在特定业务场景下可能具有情感倾向。

针对以上大规模文本分类存在的问题,传统解决方案使用人工特征工程+浅层分类模型方法,将文本分类问题拆分成特征工程和分类器两部分,见图1。

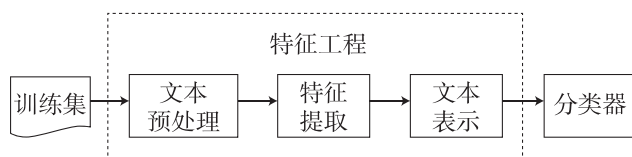


图1 传统机器学习文本分类方法图

特征工程:文本特征工程分为文本预处理、特征提取、文本表示三个部分,目的是把文本转换成计算机可理解的格式,并封装足够用于分类的信息,即很强的特征表达能力。

分类器:大部分传统浅层机器学习方法都可以应用于文本分类领域,比如朴素贝叶斯分类算法(Naïve Bayes)、KNN、SVM、最大熵和神经网络等。

基于传统分类模型的文本分类方法主要集中在特征提取和选择上,常用的方法有 TF-IDF、词频、文档频次、N-Gram、互信息等。但是文本表示是高纬度、高稀疏的,特征表达能力很弱,神经网络不擅长对此类数据的处理,而且还需要使用人工特征工程,成本很高。

2.2.2 基于深度学习的文本分类方法

近年来,随着深度学习的不断发展,学者们将文本分类的研究重点转移到了基于人工神经网络的分类模型上。人工神经网络简称神经网络,是模拟生物神经网络进行信息处理的一种数学模型。人工神经网络由众多可调连接权值的神经元连接而成,它的参数学习基于误差反向传播(Back-Propagation, BP)算法,具有很强的非线性映射能力。利用神经网络处理文本分类的优势之一是无需花费大量的时间在特征提取与选择上,将词的分布式表示作为特征输入到网络中,神经网络可以自动抽取对文本分类有价值的信息,通常这些信息是经过卷积、点乘、非线性函数、矩阵相乘等操作得到,且所得信息高度编码不易被解读。

与传统机器学习不同,深度学习既提供特征提取功能,也可以完成分类功能,并且可解决传统机器学习方法存在的问题,具体实现流程包含模型预训练和迁移学习两个阶段,如图2所示。

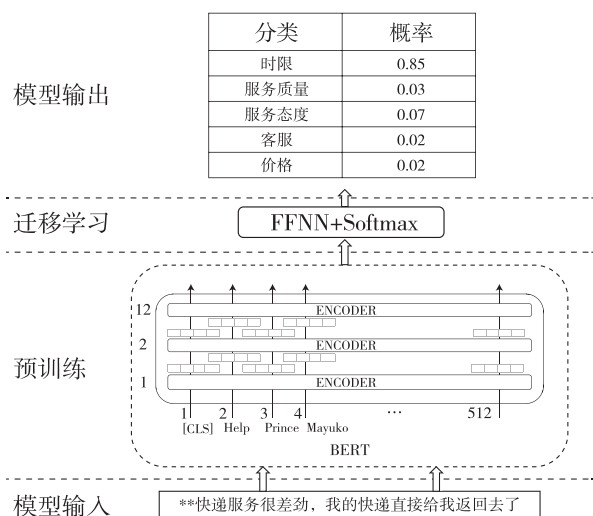


图2 深度学习文本分类模型两阶段训练方法图

阶段一：模型预训练。预训练通过设计好一个网络结构来做语言模型任务，把大量语言学知识抽取出来编码到网络结构中，这些先验的语言学特征会对后续文本分类任务有特征补充作用，集成尽量通用的语言学知识能够加强模型的泛化能力。具体网络结构方面，包括基于全连接网络(word2vec、FastText)、卷积神经网络(TextCNN)、循环神经网络(TextRNN、TextAttBiRNN、TextBiRNN、RCNNVariant)、循环卷积神经网络(RCNNVariant)、

注意力模型(BERT)等方法,最终实现并行计算,同时保持长距离上下文特征提取能力,有效获得基于上下文的词向量表示。

以TextCNN模型为例,进行模型结构的介绍:TextCNN可以看作是n-grams的表现形式,例如三种不同的feature size的卷积核对应了3-gram、4-gram和5-gram,整体模型结构如图3所示,先用不同尺寸(3, 4, 5)的卷积核去提取特征,再进行最大池化,最后将不同尺寸的卷积核提取的特征拼接在一起作为输入到softmax中的特征向量。

输入层:文本矩阵。对于文本中的词语可先构建大的词库集,用word2vec进行训练,获取得到每个词语的词向量,所有词向量拼接能够得到对应的文本矩阵。

卷积层:使用不同的卷积核,卷积核的宽度和词向量的长度一致,每个卷积核获得一列feature map。

Max-pooling:每个feature map通过maxpooling都会得到一个特征值,这个操作也使得TextCNN能处理不同长度的文本。

全连接层:该层的输入为池化操作后形成的一维向量,经过激活函数输出,再加上Dropout层防止过拟合,并在全连接层上添加L2正则化参数。

Softmax:将全连接层的输出使用softmax函数,获取文本分到不同类别的概率。

阶段二:迁移学习。迁移学习是在预训练的基础上获得的词向量,表示加一层softmax网络,然后在已经打好标签的训练数据上重新训练,利用已知的网络结构和已知的网络参数,修改output层为文本分类层,微调最后一层前的若干层的参数。

基于深度学习的方法,可解决大规模文本分类问题中的基于上下文的文本语义表示问题,再利用CNN/RNN/Attention等网络结构自动获取特征表达能力以及迁移学习强大的泛化能力,去掉繁杂的人工特征工程,形成端到端的解决方案,有效攻克了分词难、缺乏标准完整的情感词库、否定词表达多变、多标签分类难等难点。

3 研究过程及结果

3.1 研究方法及过程

为进一步分析公众对邮政及竞品的寄递服务

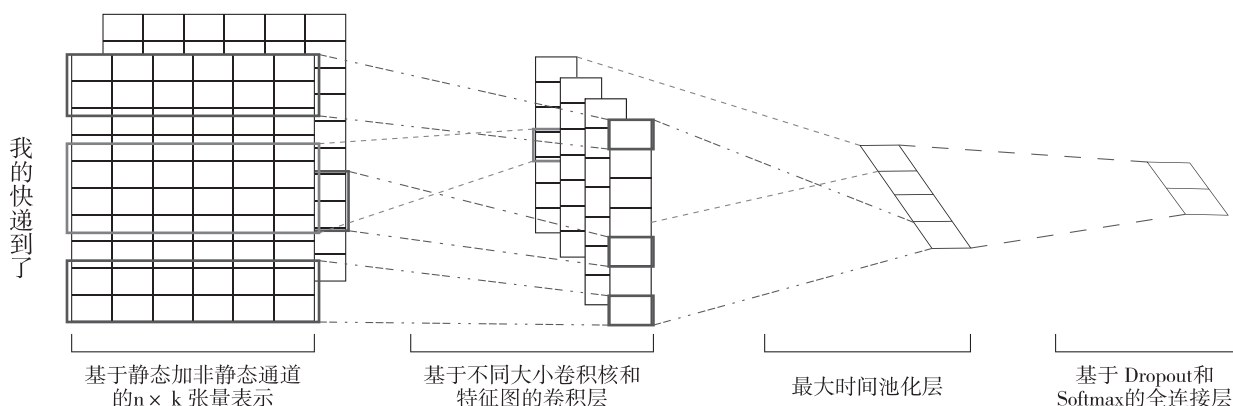


图3 Text-CNN模型结构图

发表舆论关注点的集中情况，本研究研发了基于深度学习的正负面舆情分类模型、舆情关注点分类模型。正负面舆情分类模型将社交媒体平台上的客户评价文本分为正面、负面和中性评价；从用户体验的触点出发，将舆情关注点分类模型进一步分为正面5类、负面5个一级分类+22个二级分类，分类体系如图4所示。

通过自主研发，应用基于迁移学习的深度学习方法，解决了大规模文本分类问题中基于上下文的语义表示问题；利用CNN/RNN/Attention等网络结构自动获取特征表达能力，省去了繁杂的人工特征工程，形成端到端的解决方案；通过采用迁移学习的方法，减少了对标签数据的依赖，极大提高了模型的稳定性和可泛化性。对比传统的浅层机器学习和统计学方法，取得了更高的文本分类准确率，

平均准确率达到94.77%。

3.2 研究结果

3.2.1 投诉量分析

2020年9~11月，邮政、竞品1和竞品2的投诉量均有显著增长。11月竞品1投诉量环比增长260.6%，竞品2环比增长30.2%，邮政环比增长67.5%，邮政客户投诉增长率低于竞品1（如图5所示）。

3.2.2 投诉时间分析

由于2020年“双11”大促较往年提前启动，邮政、竞品1、竞品2在11月投诉量均在11月5号达到峰值，而不是11月11号之后。结合11月时限对标结果，11月全程时限也在月初最慢；从11月整体来看，相比竞品1、竞品2，邮政投诉量波动较小，详见图6。

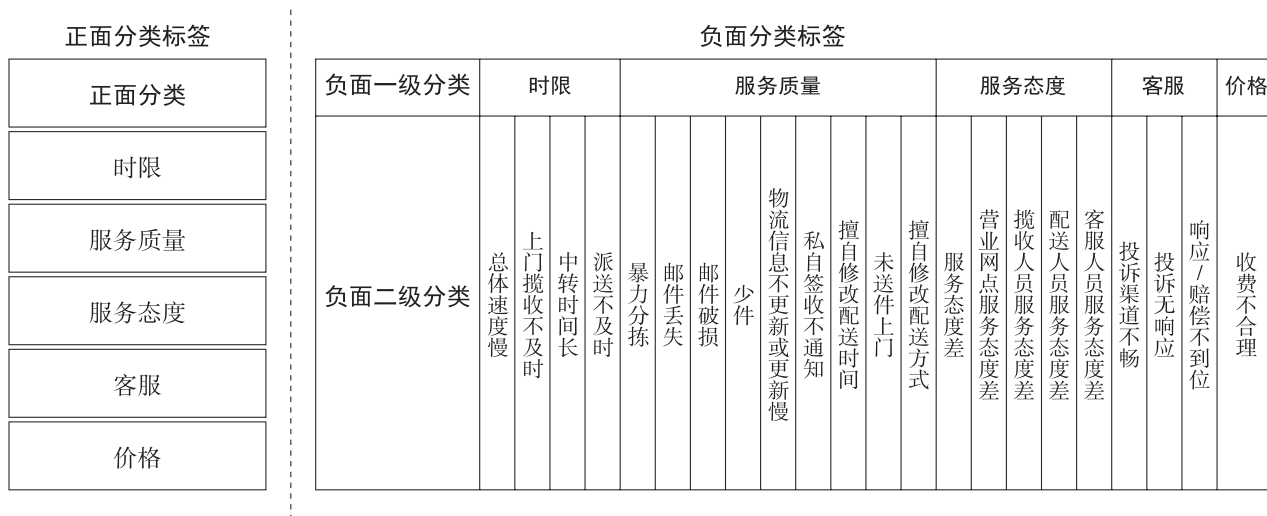


图4 舆情关注点分类体系示意图

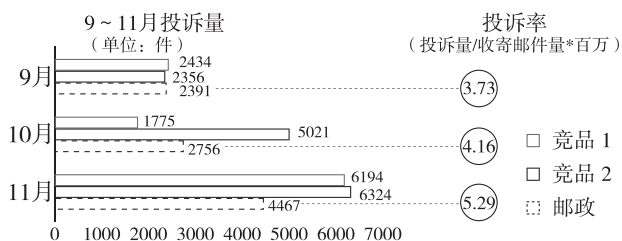


图5 2020年“双11”期间投诉量增长图

3.2.3 响应速度分析

与竞争对手比，邮政投诉响应速度明显快于竞品2，但与竞品1还有差距，比竞品1低14.7%，详见图7。

3.2.4 投诉原因分析

从社交媒体平台上的投诉内容来看，客户对邮政投诉的高频热词依次为“慢”（占比为9.85%）、“不派送”（占比为5.11%）、“延误”（占比为4.37%）。

将2020年11月客户发表的投诉内容文本输入到深度学习文本分类模型中，从一级原因来看，邮

政、竞品2和竞品1投诉量均集中在“时限”问题，占比均在60%以上，详见图8。

2020年11月，在二级原因上，各家快递投诉的主要原因均为“总体速度慢”，其次为“派送不及时”“投诉渠道不畅”，详见图9。

3.2.5 舆情传播路径分析

为研究邮政微博传播的关键用户，采用传播关系图的方法进一步研究，某条热门微博传播覆盖的用户有20473人，“大V”的用户有1038人，占比为6.38%，所贡献的转发量占比达到了43%，其中贡献率排名最高的三位“大V”为@我是**、@管鑫***和@马库**。其中@管鑫***同时出现在邮政负面和竞品3正面的主要转发节点上；此外，发现“大V”用户@颜****和@快递****，在竞品1正面舆情和邮政负面微博中均有较高的转发量。以上属于需要重点关注的“大V”用户。

3.2.6 用户特征分析

为研究热门微博覆盖的用户特征，对微博的

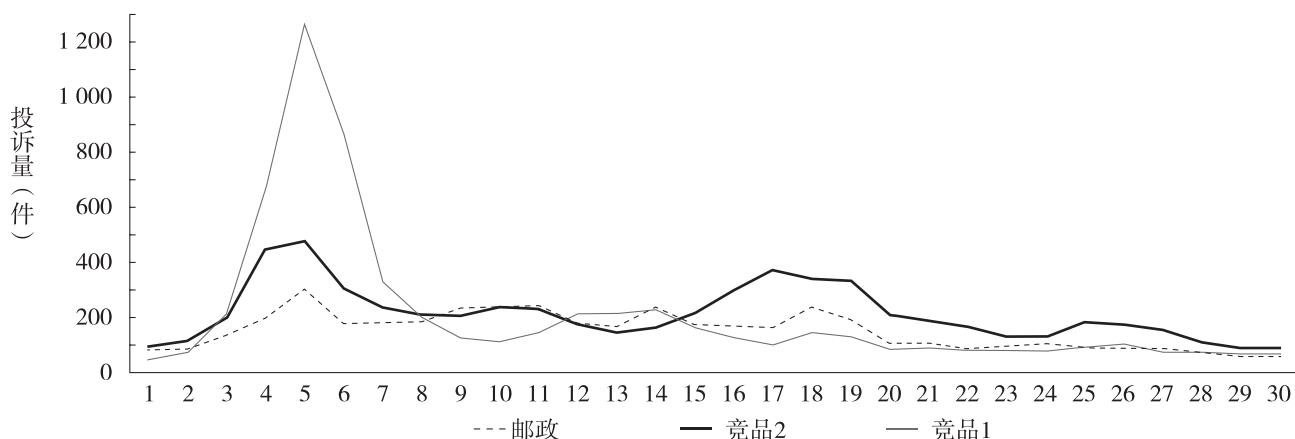


图6 2020年11月每日投诉量分布图

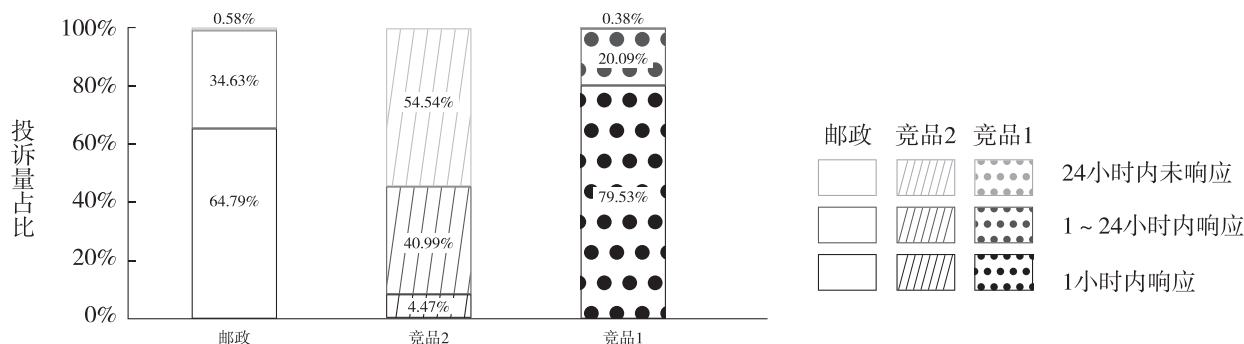


图7 “黑猫平台”投诉响应速度分布图

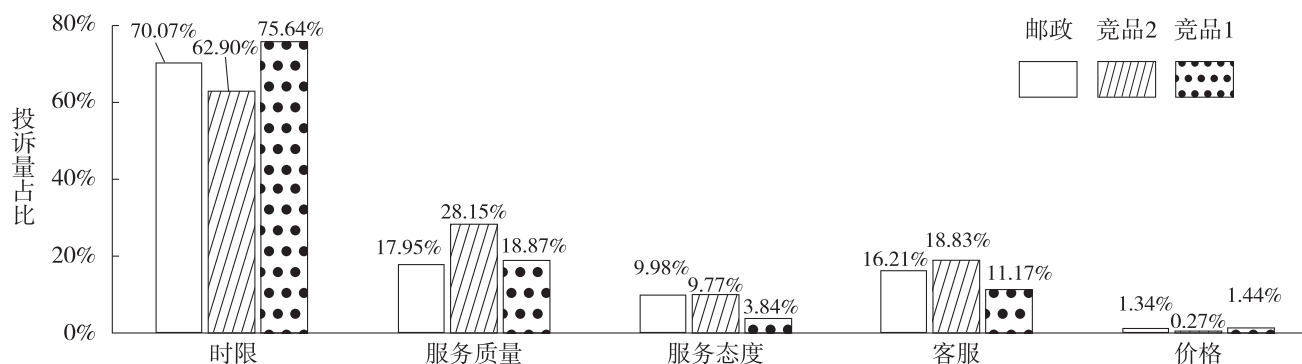


图8 投诉问题一级原因分类对标情况

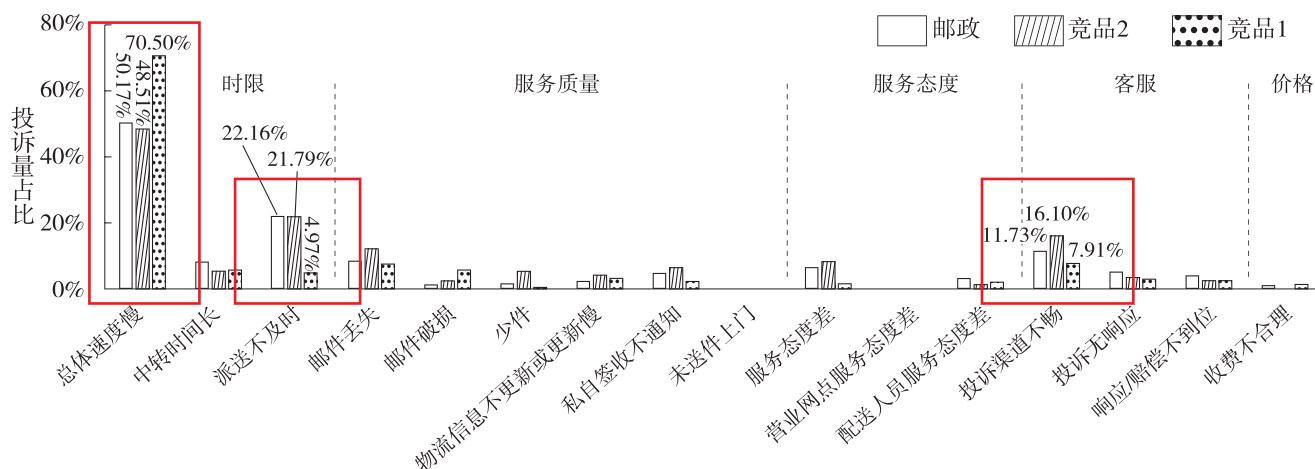


图9 投诉问题二级原因分类对标情况

转发用户进行了进一步分析。热门舆情微博转发用户中，女性用户居多，占比高达74.99%；从用户地域分布看，广东、北京、上海的用户明显活跃度较高，详见图10。

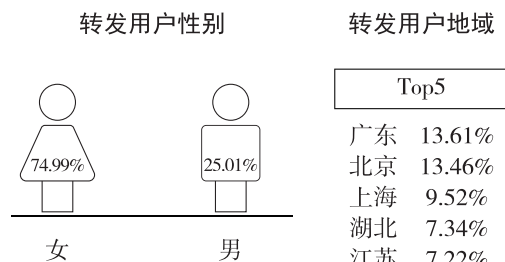


图10 用户特征分析图

4 结语

该研究为业务部门及时掌握社交媒体平台的公众舆论导向、提升智能化舆情信息管理起到积极的促进作用，有效预防及减轻企业爆发舆情危机的风险，同时通过从客户舆论的视角审视邮政服

务问题、短板和差距，为有针对性地制定改进措施，促进服务品质的持续提升提供了数据支撑。

参 考 文 献

- [1] KIM Y. Convolutional neural networks for sentence classification [J]. Eprint Arxiv, 2014
- [2] 王国薇, 黄浩. 基于深度学习的文本分类方法研究 [D]. 乌鲁木齐: 新疆大学, 2019
- [3] 刘婷婷, 朱文东, 刘广一. 基于深度学习的文本分类研究进展 [J]. 电力信息与通信技术, 2018 (3)
- [4] 黄晓斌, 赵超. 文本挖掘在网络舆情信息分析中的应用 [J]. 情报科学, 2009 (1)
- [5] 万源. 基于语义统计分析的网络舆情挖掘技术研究 [D]. 武汉: 武汉理工大学, 2012
- [6] 郝超, 袁杭萍, 孙毅, 等. 多标签文本分类研究进展 [J]. 计算机工程与应用, 2021 (10)
- [7] 张永成, 王怀彬. Python 自然语言处理方法在文本情感分析中的应用 [J]. 电脑知识与技术, 2020 (36)