居民社区在线聊天热点话题的情感分析研究

蔡云戈,范永胜,冯　骥
(重庆师范大学 计算机与信息科学学院,重庆 401331)

摘　要:为了更好地获悉社区居民在网络中所反映的民生问题,把握亟需关注的热点,基于便捷高效、沟通互动性强的在线聊天数据,提出了一种基于情感与热点话题的综合分析模型。首先,采用半监督的情感标注模型与基于注意力机制的双向长短期记忆网络模型对社区相关数据进行居民情感分析;其次,通过隐狄利克雷分布主题模型对热点问题进行研究;最后,结合话题类别与情感分布进行综合分析。实验结果表明,采用半监督的情感分类模型最终分类准确率可达到 89.92%,相较于其他基线模型,取得了更好的分类效果。经卡方检验后可知热点话题与情感分布之间具有相关性,不同社区的居民关注的话题、发言的数量及发言的长度等均存在较大的差异,各社区集中讨论的时间点与其从事职业具有密切关系。这些均可为居民社区服务部门、社区治理部门及相关社会工作者的工作提供切实有效的参考依据。

关键词:居民社区;情感分析;在线聊天;热点话题;注意力机制
中图分类号:TP391.1　　**文献标识码**:A　　**文章编号**:1673-629X(2023)05-0042-07
doi:10.3969/j.issn.1673-629X.2023.05.007

Study on Sentiment Analysis of Hot Topics in Residents' Community Online Chat

CAI Yun-ge,FAN Yong-sheng,FENG Ji
(School of Computer and Information Science,Chongqing Normal University,Chongqing 401331,China)

Abstract:To better understand the livelihood issues reflected by community residents on the Internet and grasp the hot topics that need urgent attention,a comprehensive analysis model based on sentiment and hot topics is proposed by using convenient,efficient,and communicable interactive online chat data. Firstly,a semi-supervised sentiment annotation model and an attention-based bidirectional long-short term memory network model are used to analyze community-related data for resident sentiment,followed by a latent Dirichlet allocation topic model for hot issues,and finally,combining topic categories and sentiment distribution be explored. The experimental results show that the final classification accuracy of the semi-supervised sentiment classification model can reach 89.92%,which achieves better classification results than other baseline models. A chi-square test shows that there is a correlation between hot topics and the distribution of sentiment. There are significant differences in the topics of interest,the number of statements and the length of statements made by residents in different communities,and the point in time when discussions are concentrated in each community is closely related to their occupations,which can provide a valuable reference for the work of community service departments,community governance departments, and relevant social workers.

Key words:residential communities;sentiment analysis;online chat;hot topics;attention mechanism

0　引　言

截至 2021 年 12 月,第 49 次《中国互联网络发展状况统计报告》指出以 QQ 为代表的即时通信应用类用户规模达 10.07 亿,在整体网民中占比 97.5%[1]。同年六月在 STATISTA 发布的《China: reasons for using social networks on mobile phones》调查报告中指出,休闲聊天是中国网民使用社交媒体的主要原因[2]。对居民社区而言,生活社区在线交流群(如 QQ 群和微信群等在线交流工具)是居民快速便捷向物业或相关部门反映民生诉求、解决相应问题的关键场所。但存在人员多、发言门槛低及素质参差不齐等因素造成消息密度大与信息内容杂等问题,使得民生问题得不到有效关注,热点问题得不到及时解决,从而导致居民负向情绪加剧,邻里冲突、维权受阻与矛盾激化等现象[3]

收稿日期:2022-07-24　　**修回日期**:2022-11-26
基金项目:重庆师范大学(人才引进/博士启动)基金项目(17XCB008);教育部人文社会科学研究项目(18XJC880002)
作者简介:蔡云戈(1997-),女,CCF 会员(E2097G),硕士研究生,研究方向为数据挖掘和舆情分析;通信作者:范永胜,博士,副教授,研究方向为自然语言处理。

时有产生。因此,从纷繁复杂的聊天数据中快速获取并分析人们对热点问题的情感倾向成为居民、政府部门、学者等各方关注的焦点。基于此,在收集了大量的社区聊天信息的前提下,进行了一次有意义的尝试研究,该研究主要贡献如下:

(1)针对目前对于民生领域居民社区群聊关注较少的问题,该研究通过追踪搜集大量居民社区在线聊天信息,构建生活社区群聊数据集,并对数据进行了情感与热点话题的综合分析;

(2)针对在中文社区群聊领域,因涉及隐私暂无公开标注的数据集可用于情感分类模型的有效训练的问题,构建了社区领域情感词典,并结合基于注意力机制[4]的双向长短期记忆网络(Attention-based Bidirectional Long-Short Term Memory Network, Att_BiLSTM)[5]情感分类模型实现了对社区群聊文本的半监督情感倾向计算;

(3)结合生活社区热点话题验证了话题与情感间的相关性,并举例展示了两类社区居民关注的话题、发言数量和长度等特征,发现各社区集中讨论的时间点与其从事职业具有密切关系。相关分析结果可为有关管理部门及人员提供切实有效的参考依据。

1　研究现状

针对这样的在线聊天文本,目前已有研究者们从话题与情感两方面开展了相应研究。

在国外,Pellert 等人[6]使用奥地利不同数据源在 Twitter 及学生聊天群等多平台中搜集新冠疫情期间相关内容进行情感分析,通过可视化界面展示了疫情期间各类情感变化与内容分布;Ng 等人[7]从参与度、情绪与话题讨论等五个维度出发,采用 MPQA 词典检测群内人员对舆论和权威信息的态度,分析新加坡某群聊中人们在新冠疫情期间对错误消息传播的反应;Saha 等人[8]在全球最受欢迎的即时通信应用 WhatsApp 中,分析了数千个讨论印度政治的群组有关恐怖言论的用词特性、主题分布及传播特征,对恐怖信息进行了有效的舆情检测,在一定程度避免了恐怖言论的恶性传播。

在国内,张大勇等人[9]以微信群为例分析了群体互动行为特征,结果表明带有情绪诱导和相关利益引导的标题可引起更多用户的互动;汪鸿沁泠等人[10]从话题交流强度、成员活跃度及话轮密度三个维度对群聊文本的话题进行强度计算与演化分析,得到群聊话题演化的生命周期规律及热点结构;吴旭等人[11]综合话题序列、群聊内容等因素提出了多策略话题检测模型,扩大了话题检测所能应对消息类型的广度,提升了舆情分析效率。

国内外学者对于社区群聊展开的研究进行了广泛的探索,但对于民生领域的关注相对较少。为此,笔者通过对社区群聊的半监督情感倾向计算,结合相关性及可视化分析方法得到社区居民对不同话题的情感倾向及发言特征,以期提升管理者对民生诉求的关注度,辅助相关人员高效管理社区,提高社区居住幸福感[12]。

2　数据处理流程及相关模型介绍

2.1　数据处理流程

社区群聊内因发言人群素质差异大、口语化严重、话题跳跃度高等特性,相较于普通文本的数据处理流程,需要针对具体细节做出相应修改,如新增生活社区情感词典等,以提高后续对数据的处理效率及情感分类准确性。

为此,构建了如图 1 所示的社区情感与热点话题关联性分析处理技术流图,其大致可分为 3 个主要过程。

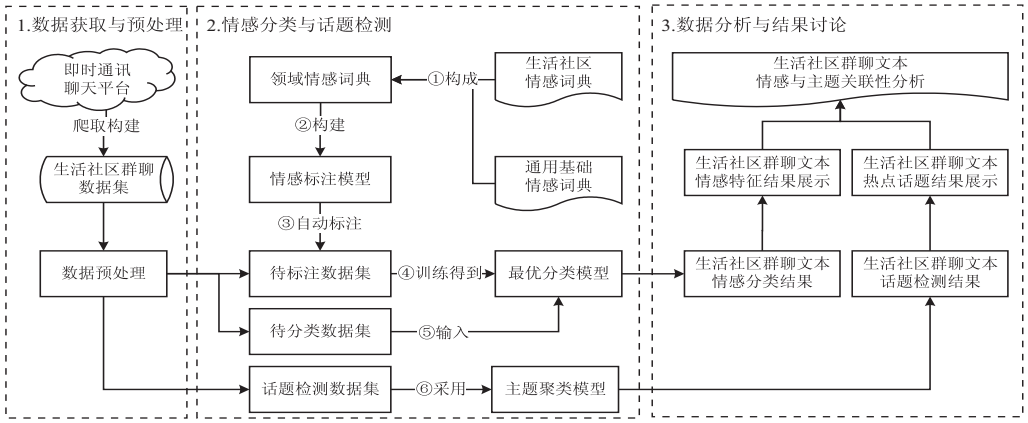


图 1　社区情感与热点话题关联性分析处理技术流图

(1)数据获取与预处理。

在数据获取阶段,通过即时通讯软件的相关聊天平台收集了 2020 年 10 月至 2022 年 5 月期间几十个不同平台的生活社区群聊数据共 1 038 604 条,从中提

取聊天成员、聊天时间和聊天内容等信息构建生活社区群聊原始数据集。在数据预处理阶段,经过表 1 中的相应方式对异常数据进行处理,最终得到有效数据 727 023 条。使用全部数据构建话题检测数据,从有效数据中随机抽取 10% 构建待标注数据集,剩余 90% 组成待分类数据集。

表 1 异常数据处理

数据位置	异常情况	数据范例	处理方法
发布时间	空值		删除该条数据
聊天发布者	空值		统一填充
聊天内容	空值		删除该条数据
	无效值	广告链接等无实际内容与意义的信息	删除该条数据
	刷屏值	同一人同一时刻重复发布相同信息	删除重复值
	情感符号	[大笑]、[撇嘴]	转为文字内容

(2)情感分类与话题检测。

对于情感分类而言,具体步骤如下:①基于 TF-IDF 与 SO-PMI 算法[13]对有效数据进行种子情感词获取与生活社区情感词典的构建;②结合通用情感词典,构建领域情感词典,并依据词本身是否积极以及其前后所使用的修饰词来决定其加权的正负和权重这一规则构建情感标注模型,完成对待标注数据集的情感标注;③采用经标注后的数据对各分类模型进行训练与评估,依据实验结果验证标注数据质量,并选择最优分类模型完成剩余待分类数据集的情感分类。

对于话题检测而言,具体步骤如下:①在话题检测数据集中按照社区群类型分别通过隐狄利克雷分布(Latent Dirichlet Allocation,LDA)[14]主题聚类模型进行话题检测;②结合可视化结果分析得到最优话题数和相关特征词。

(3)数据分析与结果讨论。

依据上述情感分类与话题检测结果对相应特征进行可视化分析,再结合二者进行相关性分析[15]。

2.2 相关模型简介

2.2.1 分类模型简介

该文分别采用如下 6 种分类模型进行了对比实验,各模型原理及优缺点总结如表 2 所示。

表 2 分类模型

模型名称	原理	优点	缺点
Random Forest[16]	以决策树为基础,由 Bagging 集成构成的传统机器学习分类算法。通过结合多个弱分类器的投票结果改善分类器的泛化性,最终分类结果为投票最高的类别	(1)各学习器可并行训练,训练速度快; (2)引入随机特征选择,模型抗噪能力强,不易过拟合	(1)模型可解释较差; (2)无法控制模型内部的运行
XGBoost	以决策树为基础,由 Boosting 集成构成的传统机器学习分类算法。通过组合多个弱分类器构成一个强分类器,从而得到更优的分类结果	(1)借鉴 RandomForest 支持列抽样,减少计算量,降低过拟合; (2)可自定义损失函数调整模型,获得更好的训练效果	(1)参数多;空间复杂度高; (2)对于高维特征数据,处理能力弱于深度学习算法
TextCNN[17]	通过多个不同尺寸的窗口提取句子中的关键信息来捕捉局部相关性,将词向量经一层卷积和最大池化后通过激活函数得到分类结果	(1)模型结构简单 (2)训练速度快	受限于固定窗口的大小,无法建模获得更长序列信息
LSTM[18]	以循环神经网络(RNN)为基础,加入输入门、遗忘门、输出门和内部记忆单元来改变神经元运算公式以提高对信息的筛选能力,由此获得更优分类结果	(1)具有长时记忆功能; (2)可选择性保存有用信息、遗忘无用信息,解决长时记忆中信息丢失问题	(1)并行处理存在劣势; (2)无法从后往前编码信息
BiLSTM	通过组合前向和后向 LSTM,解决了 LSTM 无法从后往前编码信息的问题,实现对上下文信息的双向考虑,进一步提升分类准确性	(1)可同时捕获前向与后向相关语义信息; (2)有利于对语义信息进行双向理解	各模块输出信息对结果的影响程度相同,无法凸显关键信息对分类结果的影响性
Att_BiLSTM	在 BiLSTM 模型的基础上加入注意力机制对原有的 Encode-Decode 框架进行改进,将双向 LSTM 的输出结果输入至注意力机制中,使其能够自动学习与计算输入数据对输出数据的贡献大小,使模型从大量信息中捕获关键的句子信息,实现更好的分类效果	(1)打破编码时对向量的长度限制; (2)可按权给予各信息不同的关注度,突出关键信息对最终分类结果的影响	(1)模型内部结构复杂; (2)训练时间长

其中 Att_BiLSTM 模型结构如图 2 所示。

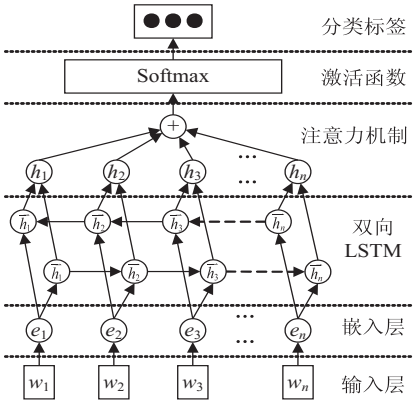


图 2　Att_BiLSTM 模型结构

2.2.2　主题聚类模型简介

该文采用 LDA 主题模型识别不同社区间热点话题的差异。LDA 主题模型属于无监督学习模型,由文档、主题和词语构成的三层贝叶斯概率模型组成,通过概率统计方法对文档中选出的关键词语进行主题归纳,具有结构简单、训练速度快且聚类效果直观等优点。为获得最佳聚类效果,该文通过多次迭代,动态调整参数和观测 pyLDAvis 可视化效果确定最佳话题聚类数并挖掘热点话题特征词。

2.2.3　评价指标

(1)分类模型评价指标。

该文采用准确率(Accuracy)、精确率(Precision)、召回率(Recall)以及调和平均值(F1)的加权平均(Weighted average)计算方法作为模型评价指标,各指标计算方法如式(1)至式(4)所示。

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}} \tag{1}$$

$$P = \frac{1}{n}\sum_{i=1}^{n} W_i P_i, \quad P_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i} \tag{2}$$

$$R = \frac{1}{n}\sum_{i=1}^{n} W_i R_i, \quad R_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i} \tag{3}$$

$$\text{F1} = \frac{1}{n}\sum_{i=1}^{n} W_i \frac{2 \times P_i \times R_i}{P_i + R_i} \tag{4}$$

其中,TP 为正例预测为正例的样本数;TN 为反例预测为反例的样本数;FP 为反例预测为正例的样本数;FN 为正例预测为反例的样本数;n 代表类别数;i 代表第 i 类样本;W_i 代表第 i 类样本的总数在总样本中的占比权重。

(2)相关性评价指标。

卡方检验是以卡方分布为基础的一种检验方法,经常用于检测多个定类变量间是否存在相关性。在检验之前需做出零假设即假设两个变量呈统计独立性。其次对两定类变量建立列联表,若列联表共有 r 行 c 列,则样本自由度 df 与各字段的期望频数 E 如式(5)和式(6)所示。

$$\text{df} = (r - 1)(c - 1) \tag{5}$$

$$E_{i,j} = \frac{(\sum_{n_c=1}^{c} O_{i,n_c}) \bullet (\sum_{n_r=1}^{r} O_{n_r,j})}{N} \tag{6}$$

其中,N 代表样本总个数,O 代表指定位置下的实际观测值,i 与 j 分别代表第 i 行与第 j 列,n_c 与 n_r 分别用于遍历第 i 行所有列的值与遍历第 j 列下所有行的值。

接着,依据式(7)计算各字段皮尔逊卡方值 χ^2,当两定类变量间关联程度越强时,皮尔逊卡方值 χ^2 也会越大。

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}} \tag{7}$$

最后,通过卡方值和自由度查表得出卡方分配右尾概率 P 是否位于拒绝域内,可判断显著性水平。当 $P < 0.05$ 时,得到拒绝原假设,两分类变量间存在显著性差异的结论,即证明两个分类变量间具有相关性。

3　实验过程

3.1　实验环境

为验证算法有效性,基于如下环境开展实验。(1)硬件环境:11th Gen Intel(R) Core(TM) i5-11400F CPU(频率为 2.60 GHz),显卡为 NVIDIA GeForce RTX 3080 GPU,内存为 32.0 GB;(2)开发环境:实验均在 Windows 10 操作系统、PyTorch 1.11.0、CUDA 11.3 和 Python3.8 环境下运行。

3.2　实验数据

经 2.1 小节中相关预处理操作后,通过情感标注模型对待标注数据集进行数据标注,可得到情感分类模型训练数据,如表 3 所示。

表 3　情感分类模型部分训练数据示例

原序号	聊天时间	聊天者	聊天内容	情感标注
8622	2020-12-31 20:55:20	4791 * *06	这不像一个 2000 年以后的小区,倒是颇有点上世纪 70 年代老旧小区乱搭乱建的感觉,这管道走的……	2
9031	2021-01-04 17:05:13	7870 * *715	还有 8 号线,6 号线,3 号线等都分了好几期修建	0
17087	2021-04-05 20:51:52	3037 * *106	光谷之星 26000 那套是毛坯还是精装	1
23864	2021-04-27 12:45:34	4207 * *788	好的,工人来了发信息。谢谢	1
25823	2021-05-12 13:57:35	3782 * *555	89 折?	0
……	……	……	……	……
80721	2022-03-31 9:02:12	7827 * *14	咋找到的? 跑回来的?	2

3.3 实验结果

对情感分类数据集中 72 703 条数据按照 6∶2∶2 的比例划分为训练集、验证集与测试集。并采用 Att_BiLSTM 等多个分类模型进行实验后,各模型在测试集上的实验结果如表 4 所示。实验结果证明了 Att_BiLSTM 模型在此领域各项指标均表现最优。

表 4 情感分类模型实验结果对比

Model	Accuracy	*P*	*R*	F1
RandomForest	64.69	66.82	64.69	63.46
XGBoost	70.63	70.72	70.63	70.48
TextCNN	73.92	74.03	72.95	73.41
LSTM	86.50	79.63	88.89	83.84
BiLSTM	87.59	88.89	88.89	88.89
Att_BiLSTM	89.92	91.67	88.89	89.07

为了更好展现生活社区中居民的情感趋势与热点话题之间的联系,选取 2021 年期间两个典型社区 A(工业区)与 B(文化区)的相关数据进行情感特征分析与热点话题检测。

4 结果分析

4.1 发言长度与情感极性分析

经统计 2021 年 A 社区发言总数为 46 974 条,B 社区发言总数为 6 103 条,A 社区活跃度远高于 B 社区。对两社区各类情感发言数及每条文本平均长度进行统计,结果如表 5 所示。

表 5 发言数与各类情感平均文本长度统计

社区名	正向		负向		中性	
	发言数/条	平均长度/字	发言数/条	平均长度/字	发言数/条	平均长度/字
A	12 611	13.53	13 003	12.88	21 360	9.13
B	2 404	26.99	1 885	26.54	1 814	13.71

由表 5 可知在 A 社区内,中性发言占比达 45.47%,远高于正向与负向情感的发言量;而 B 社区内各类情感所对应的发言量较为平均,其中正向发言占比最高,约为 39.39%。就发言长度而言,B 社区三类情感下的平均发言长度均明显高于 A 社区。进一步对各类情感下发言条数及每条发言对应使用的词语长度进行统计分析,可得到图 3 中所示结果。

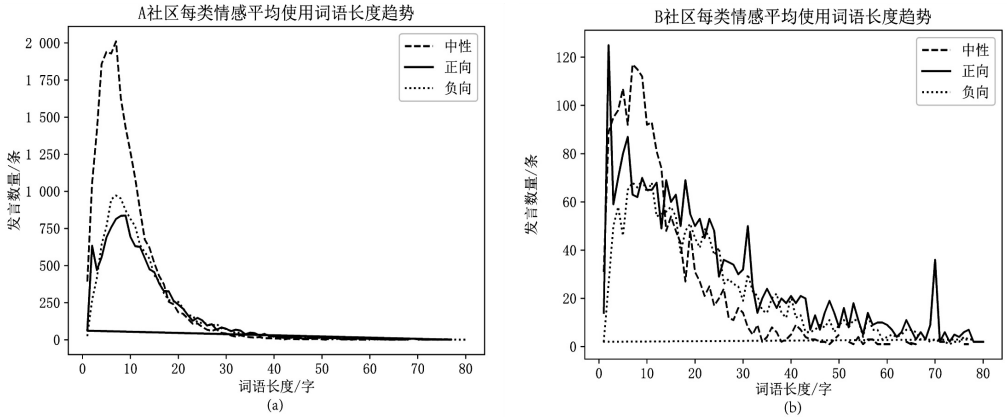


图 3 A 与 B 社区内发言词语长度与发言条数统计

由图 3 可知,多数群聊文本的词语长度集中在1 ~ 20 个汉字之间,长度越长对应的发言量越少。对于 A 社区而言,词语长度在 2 ~ 11 个汉字间时,中性情感发言量远高于正负向情感发言,在词语长度大于 20 之后,发言的情感走向就难以区分;对于 B 社区而言,当词语长度在 1 ~ 15 区间内时,表现出的情感倾向基本与 A 社区相同,但当词语长度大于 15 后,B 社区内的发言就带有明显的个人情感倾向。

4.2 发言周期性特性分析

按照以天为周期与以季度为周期对发言量进行统计分析,可得图 4 中所示结果。

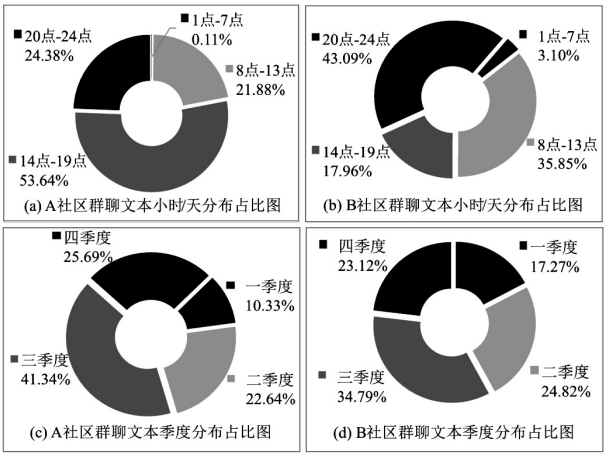


图 4 发言量/周期占比分布

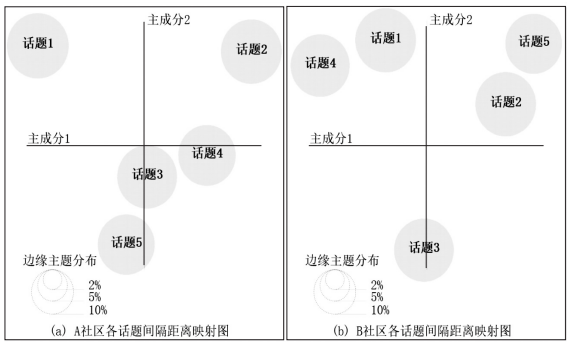
从图 4(a)与图 4(b)中可看出,两社区聊天量在一天中的时间点分布存在很大不同。虽然在 1 ~ 7 点间两者群内活跃度均为一天之中的最低点,这也符合大多数中国居民的日常作息时间。但A社区在14 ~ 19 点间的聊天量几乎是该时段 B 社区聊天量的 3 倍,而 B 社区在 20 ~ 24 点与 8 ~ 13 点间的聊天量接近于 A 社区的 2 倍,这表明社区类型和居民工作类型与聊

天时间具有紧密联系。工业区上班族大多倾向于在下午参与群内讨论,而多数学校工作者习惯于利用晚上和中午休息的时间解决生活中的问题,生活与工作间具有较明显的界限。

从图 4(c)和图 4(d)可知,在全年周期分布上两社区群均在第二、三季度即夏季和秋季活跃度最高,表明居民在该时段更关注社区事务。而第一季度是中国传统节假日安排较为集中的时段,大多数居民会利用假期放松或将投入家庭生活之中,对社区事务的关注度也因此降低,所以两社区该时段内的活跃度均为全年最低。

4.3　热点话题分析

采用 LDA 主题模型对社区文本进行话题聚类,根据迭代实验及可视化结果分析确定 A 和 B 社区最优主题聚类数为 5,话题聚类结果如图 5 所示。图中圆圈的大小代表各类话题所出现的频率,依据话题频率由高至低的顺序对话题进行编号。各圆圈间的距离采用 JSD(Jensen-Shannon Divergence)距离计算得到,可直观表达各话题间差异程度。



(a) A社区各话题间隔距离映射图　　(b) B社区各话题间隔距离映射图

图 5　A 和 B 社区话题分析可视化结果

依据上述话题聚类结果,分别对两社区各话题下的特征词进行分析,筛选各话题内排名前 12 的特征词,并定义各话题中心词,得到如表 6 所示结果。

从表 6 中可知 A 社区全年超半数的话题讨论集中在买房投资之下,讨论量占比达 55.65%;其次是对楼盘开发与学区政策等话题的讨论,这说明在经济发达的工业区,房子的户型和位置、各大楼盘的开发商管理和房屋售价等与房子相关的问题是关注的热点,因此建议管理部门从购房政策等角度入手,关注居民购房售房等问题;而生活压力相对较缓的 B 社区,接近半数的话题讨论量均与求助帮忙话题相关,其次是社区管理和事件投诉相关话题,说明 B 社区内邻里关系更加紧密和谐,居民更关注当前住房的生活质量、关心社区管理及社区内停水停电、设施破损等相关的民生问题。因此,对于此类社区,管理部门应注重社区基础设施的建设与维护,通过保障居民的衣食住行来减少居民在事件投诉话题下的讨论量。

表 6　A 和 B 社区各主题特征词展示

社区名	主题	特征词
A	话题 1(55.65%) 买房投资	房子、感觉、肯定、投资、喜欢、物业 地铁、结婚、位置、学校、户型、楼盘
	话题 2(13.43%) 楼盘开发	桃花源、光谷、好像、价格、保利、买房 开发商、地方、大道、希望、老板、韭菜
	话题 3(11.99%) 学区政策	估计、不行、孩子、学区、贷款、国家 不到、有钱、二手房、赚钱、新房、装修
	话题 4(9.72%) 周边时事	小区、中心、星河、还好、时间、朋友 商业、不想、天地、关系、男人、涨价
	话题 5(9.22%) 工作生活	房价、公司、工作、老师、工资、车位 离婚、股票、大学、套房、学生、社区
B	话题 1(45.4%) 求助帮忙	电话、雨棚、有没有、装修、校区、邻居 重师、工作人员、推荐、入住、老人、小孩
	话题 2(14.55%) 社区管理	物业、接种、时间、电梯、垃圾、投票 登记、公司、疫苗、居民、人员、回来
	话题 3(14.14%) 事件投诉	小区、学校、外墙、估计、解决、停车 漏水、影响、监控、白蚁、驾照、停车场
	话题 4(13.78%) 通知公告	老师、房子、通知、停水、朋友、装修 信息、感谢、打扰、广告、停气、转发
	主题 5(12.13%) 房屋交易	请问、房屋、社区、医院、单元、出售、 大学城、有意者、靠近、办理、精装、两厅

4.4　情感倾向与主题相关性分析

依据上述分析,统计两社区中各类主题对应的情感分布情况可得表 7 所示结果。

表 7　主题情感分布

社区名	主题	情感类别		
		正向(条)	中性(条)	负向(条)
A	买房投资	6 330	13 078	6 732
	楼盘开发	2 001	2 533	1 775
	学区政策	1 271	1 590	1 470
	周边时事	1 457	1 821	1 286
	工作生活	1 552	2 338	1 740
B	求助帮忙	1 046	1 044	681
	社区管理	307	192	241
	事件投诉	267	206	415
	通知公告	284	230	327
	房屋交易	500	142	221

由表 7 中可知,A 社区居民在“买房投资”话题下,中性情感的占比约为正向和负向情感的 2 倍,其余话题下三类情感的发言数量基本持平,中性情感略微突出,由此可看出在 A 社区内,居民参与各类话题讨论时大多持理性态度,对各类热点话题参与度较高且讨论的话题相对广泛自由;但对于 B 社区而言,各类话题下情感倾向更具有典型性,在“求助帮忙”“社区管理”和“房屋交易”话题下,B 社区居民的正向及中

性情感明显高于负向情感,但在“事件投诉”和“通知公告”话题下,居民更多展现出的是负向情感,特别是在“事件投诉”话题下,负向情感的发言量约为正向或中性情感的 2 倍。

因此,为了进一步验证话题类别与情感类别间是否存在相关性,依据表 7 对 A、B 社区分别进行卡方检验。给定原假设为话题类别与情感类别间不存在相关性,经检验后可得到表 8 所示结果。因 $P < 0.05$,依据卡方分布的规则可知在 99% 的情况下拒绝原假设,即话题与情感间具有显著性差异,可说明话题类别与情感类别间存在相关性。

表 8 A 与 B 社区话题与情感卡方检验结果

社区名	皮尔逊卡方 X^2	渐进显著性 P
A	578.838	0.000 * * *
B	366.181	0.000 * * *

注:* * * 代表 1% 的显著性水平

5 结束语

在收集了大量的社区居民在线聊天信息的基础上,结合生活社区领域情感词典,采用 Att_BiLSTM 情感分类模型实现对社区群聊的半监督情感倾向计算,经 LDA 主题模型分析生活社区热点话题后发现,热点话题与情感类别间具有相关性,如“买房投资”话题中 50% 的讨论倾向于中性情感,而“事件投诉”和“通知公告”等话题下负向情感占比是正向与中性情感的 2 倍。与此同时在参与讨论的时间分布上,居民在夏秋季对社区事物的关注高于其他时段,不同类型社区的居民一天内参与话题讨论的时间点与其从事职业具有密切关系,如工业区居民在 14 ~ 19 点之间群内讨论量占全天的 53.64%,而文化区居民该时段的聊天量占比仅为 17.96%。因此,有关部门可根据社区类型与居民讨论话题,在居民参与社区事务讨论的高峰时段对相关热点话题进行关注或介入,由此更好地获悉居民社区中所面对的民生问题,把握亟需关注的热点,为社区内创造良好沟通环境,及时解决居民诉求,提升社区居民幸福感。

参考文献:

[1] 第 49 次中国互联网络发展状况统计报告[EB/OL]. 2022-07-06. http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/.

[2] China: reasons for using social networks on mobile phones 2021 | Statista[EB/OL]. 2022-07-06. https://www.statista.com/statistics/277651/reasons-for-using-social-networks-in-china/.

[3] 邵春霞. 数字空间中的社区共同体营造路径——基于城市社区业主微信群的考察[J]. 理论与改革,2022(1):47-58.

[4] 王宇欣,方浩宇,张 伟,等. 注意力机制在情感分析中的应用研究[J]. 计算机技术与发展,2022,32(4):193-199.

[5] LI W,QI F,TANG M,et al. Bidirectional LSTM with self-attention mechanism and multi-channel features for sentiment classification[J]. Neurocomputing,2020,387:63-77.

[6] PELLERT M,LASSER J,METZLER H,et al. Dashboard of sentiment in austrian social media during COVID-19[J]. Frontiers in Big Data,2020,3:32.

[7] NG L H X,LOKE J Y. Analyzing public opinion and misinformation in a COVID-19 telegram group chat[J]. IEEE Internet Computing,2021,25(2):84-90.

[8] SAHA P,MATHEW B,GARIMELLA K,et al. “Short is the road that leads from fear to hate”: fear speech in Indian WhatsApp groups[C]//Proceedings of the web conference 2021. New York: Association for Computing Machinery,2021:1110-1121.

[9] 张大勇,许 磊,孔洪新. 社交媒体用户群体互动行为特征研究——以微信用户群分享为例[J]. 情报理论与实践,2019,42(10):97-101.

[10] 汪鸿沁泠,巴志超,李 纲. 微信群会话话题强度计算及演化分析[J]. 数据分析与知识发现,2019,3(2):33-42.

[11] 吴 旭,陈春旭. 基于多策略的群聊话题检测技术[J]. 数据分析与知识发现,2021,5(5):1-9.

[12] 陈丹引,闵学勤. 线上社区参与的邻里效应——基于社区微信群的实证分析[J]. 社会发展研究,2021,8(3):88-108.

[13] SHARMA S S,DUTTA G. SentiDraw: using star ratings of reviews to develop domain specific sentiment lexicon for polarity determination[J]. Information Processing & Management,2021,58(1):102412.

[14] 李金泽,张 鹏,王 娟,等. 基于舆情大数据的网络群体性事件动态识别模型与应对策略研究[J]. 情报科学,2022,40(5):73-83.

[15] JEONG J,KIM N. Does sentiment help requirement engineering: exploring sentiments in user comments to discover informative comments[J]. Automated Software Engineering,2021,28(2):1-26.

[16] SINGH N K,TOMAR D S,SANGAIAH A K. Sentiment analysis: a review and comparative analysis over social media[J]. Journal of Ambient Intelligence and Humanized Computing,2020,11(1):97-117.

[17] 李文亮,杨秋翔,秦 权. 多特征混合模型文本情感分析方法[J]. 计算机工程与应用,2021,57(19):205-213.

[18] LI W,ZHU L,SHI Y,et al. User reviews: sentiment analysis using lexicon integrated two-channel CNN - LSTM family models[J]. Applied Soft Computing,2020,94:106435.