

基于网络爬虫与文本挖掘的社交媒体多维度情感分析方法

孟瑞娟¹ 武茂升¹ 杨海鹰¹
MENG Ruijuan WU Maosheng YANG Haiying

摘要

社交媒体数据量大、动态性强,传统爬虫易因反爬机制失效,且难以保留文本语义结构。社交媒体文本存在口语化、缩写、表情符号等噪声,难以有效提取词性特征。对其进行情感分类,仅能判断倾向,无法量化情感强度与动态变化。为此,文章提出了一种基于网络爬虫与文本挖掘的社交媒体多维度情感分析方法。设计模块化爬虫框架,支持多源数据抓取与动态更新,采用多 URL 起始点策略,通过定义单词顺序弧保留文本架构,结合主域名服务器 IP 地址抓取,确保数据完整性与语义连贯性。引入分词、去停用词、词性标注预处理,结合 TF-IDF 权重计算,筛选高价值特征词,构建精准的词袋模型字典。通过句子权重计算生成摘要,结合情感度指标量化情感强度与信息多样性,实现情感变化的趋势分析。实验表明,该算法情感分析结果准确度达 96.72%,高效准确,能全面分析社交媒体情感。

关键词

网络爬虫; 文本挖掘; 情感分析; 情感分类; 词袋模型

doi: 10.3969/j.issn.1672-9528.2025.10.037

0 引言

随着互联网技术的飞速发展,社交媒体已成为人们表达观点、分享情感和传播信息的重要平台。微博、推特、微信朋友圈、论坛等社交媒体平台每天产生海量的用户生成内容,这些内容蕴含着丰富的情感信息,反映了公众对特定事件、产品、服务或社会现象的态度和情绪。社交媒体情感分析作为自然语言处理与数据挖掘领域的重要研究方向,旨在从非结构化文本数据中自动提取用户的情感倾向(如正面、负面、中性)及其强度,为舆情监测、市场调研、产品推荐、政策制定等提供决策支持。

现有诸多研究在社交媒体情感分析领域进行了有益探索。丛眸等人^[1]的研究聚焦于通过挖掘隐式情感语句相关事实信息作为辅助特征,对 RoBERTa 模型进行预训练,并结合 BiGRU(双向门控循环单元)与注意力池化技术来实现情感分析。然而,该方法在面临数据抓取受限的情境时,难以获取全面的辅助特征,这可能导致情感分析结果出现偏差。任楚岚等人^[2]的研究则采用了混合词嵌入、时序卷积网络等技术手段,先通过双向长短时记忆网络提取情感特征,再对特征进行加权处理并完成分类。但当数据质量较低时,注意力机制难以准确分配权重,进而导致对情感动态变化的分析不够深入。刘佳等人^[3]的研究借助长程文本信息学习情感动态变化,通过门控机制消除冗余与歧义信息,强化模态交互以分析情感。但在数据质量不佳的情况下,门控机制难以有效区分关键信息,从而对情感表达的理解产生偏差。高佳希

等人^[4]的研究先运用 TF-IDF(词频-逆文档频率)算法筛选词语,再利用多头注意力 Transformer 模型编码器提取语义信息。然而,当数据质量较低时,该方法难以全面捕捉背景信息,导致对情感细微差异的分析不够精准。这些研究为社交媒体情感分析提供了不同思路,但在数据质量受限的场景下,均面临各自的技术挑战,难以实现全面、精准的情感分析。

针对上述问题,本文基于网络爬虫与文本挖掘的社交媒体多维度情感分析方法。首先,设计模块化爬虫框架,支持从多源社交媒体平台抓取数据,采用多 URL 起始点策略,通过定义单词顺序弧来保留文本架构,并结合主域名服务器 IP 地址抓取,保障数据完整性与语义连贯性。其次,对抓取的数据进行预处理,包括分词、去停用词、词性标注等操作,再结合 TF-IDF 权重计算,筛选出高价值特征词,构建精准的词袋模型字典。然后,基于词袋模型,利用句子权重计算生成摘要,并结合情感度指标,量化情感强度与信息多样性。最后,对情感变化进行趋势分析,实现多维度情感分析。此方法有效解决了传统爬虫易因反爬机制失效的问题,且能保留文本语义结构;针对社交媒体文本的噪声问题,通过预处理技术有效提取词性特征;突破了传统情感分类仅能判断倾向的局限,能够量化情感强度与动态变化,为社交媒体情感分析提供了更全面的技术手段。并通过实验验证了该算法高效且准确。

1 社交媒体多维度情感分析方法

1.1 基于网络爬虫的社交媒体网络信息抓取

在社交媒体数据采集环节,设计模块化爬虫框架以应对复杂挑战。鉴于社交媒体平台多样且数据格式各异,该框架

1. 山西应用科技学院信息工程学院 山西太原 030062

具备强大的多源适配能力,可灵活接入不同平台接口,按需切换抓取规则,确保广泛覆盖各类数据源^[5]。为提升抓取稳定性与效率,采用多 URL 起始点策略。通过预设多个初始链接,从不同入口并行抓取,有效分散反爬虫机制的拦截压力,避免因单一入口故障导致数据采集中断,大幅增加数据获取量。在数据完整性方面,创新性地定义单词顺序弧。定义网络文本中单词顺序定义弧,其计算公式为:

$$w(a_i) = \frac{\text{freq}(a_i)}{\max_{j \in A} \text{freq}(a_j)} \quad (1)$$

式中: a_i 、 a_j 代表网络文本中表示两个单词之间关系的弧; $\text{freq}(a_i)$ 代表对应的出现频率值; $w(a_i)$ 代表对应弧的权重数值。

在抓取过程中,精准记录每个单词在文本中的先后顺序,以此完整保留原始文本架构,使抓取数据在语义表达上与原始内容高度一致。同时,结合主域名服务器 IP 地址进行抓取^[6]。当常规域名解析受阻时,通过直接访问 IP 地址绕过部分反爬限制,保障数据成功获取。此方法不仅提高了数据获取的成功率,更确保了数据的完整性与语义连贯性,为后续的多维度情感分析奠定了坚实可靠的数据基础。

1.2 基于文本挖掘的社交媒体关键信息提取

在利用网络爬虫技术抓取社交媒体网络信息后,所得到的网络数据杂乱,缺乏清晰的结构化特征,不包含可直接用于情感倾向判断的有效信息。若直接对这些原始数据进行分析,不仅效率低下,而且难以得出准确可靠的结果。

文本挖掘技术作为一种强大的数据处理工具,具有从海量文本数据中精准提取有价值信息和特征的能力,能够将复杂的文本数据转化为计算机易于理解和处理的形式,为后续的情感分析提供有力支持。基于此,本研究利用文本挖掘技术的优势,对抓取到的网络数据进行深度处理与分析。首先,对原始数据进行分词、去停用词、词性标注等预处理操作,去除噪声数据,提取关键信息。然后,根据预处理后的数据结构,构建相应的“词袋模型”,将文本数据表示为词的集合,为后续的情感分类提供基础。建立的“词袋模型”,如图 1 所示。

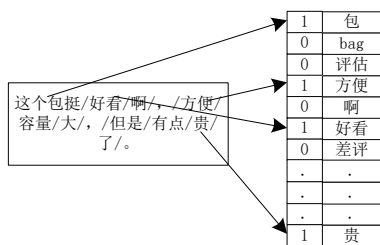


图 1 词袋模型

基于该词袋模型,进一步运用文本挖掘技术,从不同语句或图像所关联的文本中提取特征词,以此构建词袋模型的

字典。随后,针对不同的特征词及特征点进行权重赋值。

在此过程中,首先需对特征词的词频展开分析,其计算公式为:

$$f_{i,j} = \frac{t_{i,j}}{\sum_k t_{k,j}} \quad (2)$$

式中: $t_{i,j}$ 代表特征词; $f_{i,j}$ 代表指定特征词的词频。

计算其逆文件频率,分析该特征词的普遍重要性^[7],其计算公式为:

$$L_i = \log \frac{|D|}{|j: t_i \in d_j|} \quad (3)$$

式中: d_j 代表文件; L_i 代表指定特征词的逆文件频率; D 代表文档集文件总数。

根据上述计算结果,计算指定特征词的权重,其计算公式为:

$$\delta_i = f_{i,j} L_i \quad (4)$$

式中: δ_i 代表指定关键词的权重数值。

通过上述流程,得出网络信息中各关键词的权重。依据这些权重值,筛选出对分类或聚类任务最具价值的特征,并构建出“词袋模型”,以此为文本数据集。

1.3 社交媒体信息多维度情感分析

在社交媒体情感分析过程中,情感分类仅是将文本划分为正面、负面或中性等类别,这仅是对情感倾向的初步判断。实际上,还需要深入理解情感产生的原因、情感的程度以及情感的变化趋势等。由于社交媒体信息蕴含着丰富的语义、语境和隐含意义,简单的情感分类难以充分挖掘这些信息背后复杂的情感。为此,本研究在上节得到的“词袋模型”的基础上,生成社交媒体信息的摘要。根据生成的摘要,利用情感度指标^[8],对特定情感进行表达,从而实现多维度的情感分析。其计算公式为:

$$M_{p,e} = r_{p,e} \times k_p \times v_p \quad (5)$$

式中: $M_{p,e}$ 代表情感度; e 代表情感类别; p 代表反映该情感的网络信息; $r_{p,e}$ 代表网络信息模式频率; k_p 代表网络信息的逆情感频率; v_p 代表网络信息多样性。

$r_{p,e}$ 、 k_p 、 v_p 计算公式分别为:

$$r_{p,e} = \lg(f(p,e) + 1) \quad (6)$$

$$k_p = |E| / |\{e \in E : f(p,e) > 0\}| \quad (7)$$

$$v_p = \lg(\Phi(p,E)) \quad (8)$$

式中: $\Phi(p,E)$ 代表对应网络信息模式所捕捉到的情感数量。

通过上述过程,对捕获的网络信息进行全面情感分析,得到社交媒体信息中不同的情感占比^[9]。基于此占比情况,精准判断出指定信息所想表达出的情感信息,明确用户对特

定话题或事件的情感态度，如积极、消极或中立等，从而实现
对社交媒体的情感分析。

2 实验分析

2.1 设置实验环境

实验收集对于某一主题的 5 000 条社交媒体评论信息，
对其进行情感分析，并以基于语义特征提取的隐式情感分析
算法和融合 TCN 和 BiLSTM 的文本情感分析算法作为对照
组，所考虑的情感分析结果为生气、害怕、感动、开心、伤心、
惊讶 6 种模式。对设计算法设置的参数如表 1 所示。

表 1 社交媒体情感分析算法参数表

参数项	数值
词向量维度	200
空洞因子	2 ⁿ
学习率	0.03
激活函数	ReLU
扩张卷积层数	5
卷积核数	8

2.2 社交媒体网络信息情感分类能力实验结果分析

为验证本文设计算法是否能够对不同种类的社交媒体网
络信息进行情感分类，本实验应用 3 000 条信息对设计算法
进行训练，再应用剩余 2 000 条信息分为 20 组进行社交媒
体网络信息情感分类实验，分类结果为正面、负面和中性，将
分类结果与实际结果进行对比，得到每组分类结果正确的数
量，从而验证本文设计算法对社交媒体网络信息情感分类的
能力，其实验结果如图 2 所示。

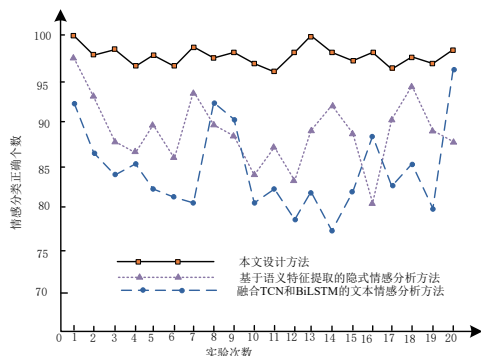


图 2 社交媒体网络信息情感分类结果

由图 2 实验结果可知，3 种算法对社交媒体网络信息的
情感分类结果（正面、负面和中性）均较为准确，且分类误
差较小。本文设计算法在每组实验中分类结果正确的数量均
在 96 以上，而对照组算法的分类正确数量均低于本文设计方
法，且数值稳定性较差。通过对比可知，本文设计算法在不
同种类情感分类任务中的表现显著优于其他两种算法。因此，
本文设计算法通过结合文本挖掘技术，有效提升了情感分类
结果的准确性和稳定性。

2.3 社交媒体情感分析准确性实验结果分析

为验证不同情感分析算法进行情感分析的准确性，本实
验对数据库中 2 000 条信息分为 20 组进行情感分析，分析结
果为生气、害怕、感动、开心、伤心、惊讶 6 种模式。其实
验结果如图 3 所示。

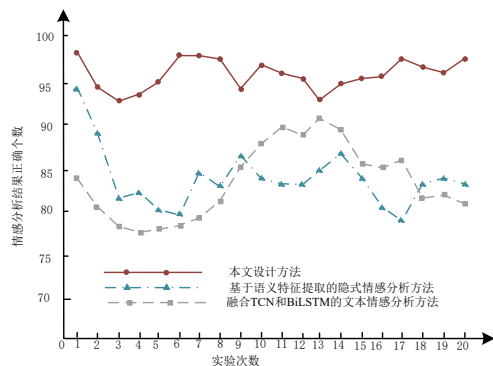


图 3 社交媒体情感分析准确性实验结果

根据图 3 实验结果所得数据计算所得，本文设计算法的
情感分析准确率为 96.72%，而对照组两种算法的准确率分别
为 86.21% 和 85.66%。通过对比可以发现，本文设计算法对
六种情感的分析结果与实际结果高度一致，尽管存在少量偏
差，但仍保持了较高的准确率。相比之下，对照组两种算法
的分析结果与实际情况存在较多不符之处。本文设计算法通
过文本挖掘技术，精准提取特征词并合理赋值，确保了关键
词提取、情感打分和摘要生成的质量，从而提升了情感分类
的准确性，为后续分析奠定了可靠的数据基础。因此，本文
设计算法在社交媒体信息情感分析中表现更为优异，且分析
结果的准确率显著提升。

3 结语

本文围绕社交媒体多维度情感分析展开深入研究，提出
了一套系统的分析方法。首先，基于网络爬虫技术构建模块
化框架，有效抓取多源社交媒体数据，并通过创新策略保障
数据完整性与语义连贯性。其次，运用文本挖掘技术对抓取
数据进行深度处理，构建“词袋模型”，提取特征词并合理
赋值，为情感分类奠定基础。在此基础上，结合情感度指标
与摘要生成技术，实现社交媒体信息的多维度情感分析，精
准判断用户情感态度。实验结果表明，本文设计算法在情感
分类准确性与稳定性上较优，在 6 种情感模式分析中展现出
高准确率。然而，本研究仍存在一定局限性，未充分考虑社
交媒体文本的实时动态性与跨语言特性，且在特征提取与情
感度计算方面仍有优化空间。未来研究将聚焦于算法优化与
拓展应用，探索实时动态情感分析方法，提升跨语言情感分
析能力，并尝试结合深度学习技术进一步挖掘社交媒体文本
中的复杂情感信息，为社交媒体情感分析提供更加全面、精
准的解决方案。

组播数据流动拓扑动态呈现系统研究

苟建国¹ 刘志伟¹ 李琳琳¹

GOU Jianguo LIU Zhiwei LI Linlin

摘要

受网络广域网方向链路带宽限制,部分业务(如视频)使用组播进行传输以减少带宽消耗。在使用中,经常出现组播数据接收不到或丢包的情况。因缺乏组播数据转发可视化工具,导致每次故障排查效率较低。为了解决上述问题,文章进行了组播数据流动拓扑动态呈现系统研究。通过远程方式获取交换机的组播信息,对组播信息进行分析,提取出组播分发树的结构以及组播接收成员的信息,最后,将这些信息拓扑动态呈现。通过检索组播信息,即可快速地知悉组播运行情况,提升网管人员组播故障处理能力和组播态势感知能力。

关键词

稀疏模式独立组播协议(PIM SM); Internet 组管理协议(IGMP); 组播; 拓扑可视化

doi: 10.3969/j.issn.1672-9528.2025.10.038

0 引言

当前使用的网管系统只能呈现网络整体拓扑结构,且需管理人员手动描绘。网络中部署的网络性能监控系统,通过使用探针分析网络镜像流量,根据报文时间戳的大小来计算报文转发路径,然而这种方式需要较多的网络探针,且在计算数据转发路径时,因强依赖报文时间戳的大小,故对网

络设备的时间同步要求较高。

同时,网络性能监控系统没有详细分析组播数据,只是单一地按照单播进行转发,依据报文时间戳计算转发路径,无法正确呈现组播报文的转发路径。Makofske 等人^[1]研究了 MHealth 系统,能够实时地呈现组播的拓扑结构;徐可夫等人^[2]开发的 MUL-Monitor 系统实现了分布式结构的组播拓扑发现方法。其研究虽然都实现了组播拓扑的可视化,但缺少组播接收成员的展示以及组播流量大小、转发接口的详细呈现,在处理组播故障时无法提供有效的信息。

1. 中国人民解放军 63626 部队 甘肃酒泉 732750

参考文献:

- [1] 丛晔,彭涛,朱蓓蓓.基于语义特征提取的隐式情感分析算法[J].吉林大学学报(理学版),2025,63(1): 107-113.
- [2] 任楚岚,仇全涛,劣思敏.融合 TCN 和 BiLSTM 的文本情感分析[J].计算机工程与设计,2024,45(7): 2090-2096.
- [3] 刘佳,宋泓,陈大鹏,等.非语言信息增强和对比学习的多模态情感分析模型[J].电子与信息学报,2024,46(8): 3372-3381.
- [4] 高佳希,黄海燕.基于 TF-IDF 和多头注意力 Transformer 模型的文本情感分析[J].华东理工大学学报(自然科学版),2024,50(1): 129-136.
- [5] 郭续,买日旦·吾守尔,古兰拜尔·吐尔洪.基于多模态融合的情感分析算法研究综述[J].计算机工程与应用,2024,60(2): 1-18.
- [6] 张昱,张海军,刘雅情,等.基于双向掩码注意力机制的多模态情感分析[J].数据分析与知识发现,2023,7(4): 46-55.

- [7] 朱亚军,次曲,拥措.基于 SVM 算法的藏文微博情感分析研究[J].计算机仿真,2022,39(8): 226-229.
- [8] 彭佳玲,周茂林,杨青.公众对上门护理服务的态度和关注点:基于网络爬虫的文本挖掘[J].护理学杂志,2023,38(5): 110-113.
- [9] 李书星,胡慧君,刘茂福.基于语感一致性的社交媒体图文情感分析[J].中国科技论文,2023,18(3): 322-329.

【作者简介】

孟瑞娟(1985—),女,山西太原人,硕士研究生,高级工程师,研究方向:大数据管理与应用、教育技术。

武茂升(1989—),男,山西太原人,硕士研究生,讲师,研究方向:大数据管理与应用、软件工程。

杨海鹰(1982—),女,山西太原人,硕士研究生,中级工程师,研究方向:计算机科学与技术、软件工程。

(收稿日期:2025-04-08 修回日期:2025-09-23)