基于机器学习的政务微博情感分析模型设计

张财[1,2]，马自强[1,2*]，闫博[1,2]

（1. 宁夏大学信息工程学院，宁夏 银川 750021；2. 宁夏大数据与人工智能省部共建协同创新中心，宁夏 银川 750021）

摘　要：针对政务微博评论杂乱、审核困难的问题，提出一种基于机器学习的政务微博情感分析模型。该模型能够量化分析政务微博中的情感，为自动审核提供有效依据。以 2022 年北京冬奥会和中国足协的微博为例，首先扩展与案例相关的词汇，并进行数据清洗和文本特征表示；然后采用机器学习模型进行情感倾向判断，并结合大连理工大学中文情感词汇文本计算情感强度。分别采用基于词袋模型和 Word2vec 模型的决策树、朴素贝叶斯和支持向量机模型进行预测，并对它们的性能进行对比评估。实验结果表明，在基于 Word2vec 的支持向量机模型下，情感分类的准确率达到 84.3%，这表明所提模型在预测政务微博情感方面具有有效性，可应用于政务微博自动审核任务。

关键词：机器学习；政务微博；情感强度；情感分析；情感分类

中图分类号：TP391　　**文献标志码：**A　　**DOI：**10.19678/j.issn.1000-3428.0068530

Design of a Machine Learning-Based Sentiment Analysis Model for Government Weibo

ZHANG Cai[1,2], MA Ziqiang[1,2*], YAN Bo[1,2]

(1. School of Information Engineering, Ningxia University, Yinchuan 750021, Ningxia, China;
2. Collaborative Innovation Center for Ningxia Big Data and Artificial Intelligence Co-Founded by Ningxia Municipality and Ministry of Education, Yinchuan 750021, Ningxia, China)

【Abstract】A machine learning-based sentiment analysis model for government Weibo is proposed to address the challenges posed by cluttered comments and subjective reviews. This model quantitatively analyzes sentiments on government Weibo, providing a reliable foundation for automatic reviews. Using the Weibo of the 2022 Beijing Winter Olympics and the Chinese Football Association as case studies, the methodology begins with the expansion of relevant vocabulary, followed by data cleaning and text feature representation. Subsequently, machine learning models are employed to assess emotional tendencies, and the Chinese sentiment lexicon from the Dalian University of Technology is utilized to calculate emotional intensity. This study employs decision trees, Naïve Bayes, and Support Vector Machine (SVM) models, incorporating both bag-of-words and Word2vec models for sentiment prediction and performance comparison. The experimental results indicate that the SVM model using Word2vec achieves an accuracy of 84.3% in sentiment classification. This demonstrates the effectiveness of the proposed model in predicting sentiments on government Weibo, indicating its potential for automatic review tasks.

【Key words】machine learning; government Weibo; sentiment intensity; sentiment analysis; sentiment classification

0　引言

根据第 51 次《中国互联网络发展状况统计报告》显示，截至 2022 年 12 月，我国网民规模达 10.67 亿，互联网普及率达 75.6%，人均每周上网时长 26.7 h[1]。10 亿用户接入互联网，形成了全球最大的数字社会。同时，随着网络舆论的快速发展，政府舆情管理面临着越来越大的挑战。互联网上的言论呈现出多样化、匿名化、分散化的特点，舆情变化速度快，且具有极大的不确定性，传统的舆情管理手段已经难以满足现代社会的需求。传统政府舆情管理系统的人工审核具有不唯一性，收支不成正比，且存在一定误差等问题，导致政府的传统舆情管理系统不再适应当前舆论发展的速度，因此，需要更加科技化和智能化的手段来有效应对舆情管理挑战。

文献[2]设计了一种融合情感词权重的情感倾向计算方法，并通过分析外汇新闻中影响情感强度的特征词，实现了最优权重组合下的外汇新闻情感强度计算。但是，他们所做的实验仅结合了外汇新闻领域词汇，对政务方面的新闻适应性不高。文

收稿日期：2023-10-09　　**修回日期：**2024-01-22

基金项目：宁夏回族自治区重点研发计划一般项目(2022BDE03008)；宁夏回族自治区重点研发计划引才专项(2021BEB04047)；宁夏自然科学基金一般项目(2021AAC03078)；国家社会科学基金项目(西部项目)(20XXW009)。

通信作者 E-mail：maziqiang@nxu.edu.cn

献[3]提出了一种基于 PAD 模型的网络口碑情感强度测度模型,用于情感倾向的测量。然而,该研究的情感词典样本量不足,而且在政务新闻的情感识别方面存在一定的局限性。文献[4]提出商品评论分析策略,实现对情感倾向的分析,并将文本的情感强度进行量化。然而,该研究的情感词典也存在一定程度的不足,特别是对于政务新闻的情感识别准确度较低。

针对政务新闻情感识别准确度较低的问题,本文通过对网络舆情的研究,设计一种新的情感倾向和情感强度测量模型,采用更加丰富的情感词典和多种文本特征表示方法及机器学习模型,以提高情感倾向和情感强度测量的准确度和可靠性。同时,本模型面向政务新闻,在政府对于网络舆论的筛选排查和管控等方面具有重要意义,进而解决政府舆情管理手段滞后、人工审核不足等问题。本研究聚焦于政务微博评论的情感分析,旨在降低内容审核的复杂性。为此,本文开发了一种融合机器学习的政务微博情感分析框架。此框架用于识别新浪微博中评论的情感倾向,为政务微博的内容管理提供支持。以 2022 年北京冬奥会和中国足协的相关微博评论作为研究案例,通过扩展相关词汇库并执行数据预处理步骤,包括文本清洗、分词和停用词去除,以准备数据集。随后,应用决策树、朴素贝叶斯、支持向量机等机器学习算法,以词袋模型和 Word2vec 模型为基础,进行情感分类的训练和测试。此外,利用大连理工大学中文情感词汇数据库计算并量化微博评论的情感强度。

采用本文所提方案,可以很好地适应政务新闻,并且在一定程度上减轻政府工作人员的工作压力,将网络舆情的情感倾向和强度进行数据化和可视化处理,从而更加直观地了解舆情的演变趋势,及时抓住网民的关注点,有针对性地开展舆情引导工作,有助于促进网络社会的和谐稳定发展。

1 相关工作

情感分析是一门涉及统计学、社会学、计算机科学等众多学科的综合性技术,旨在对文本进行深入分析和判断,以揭示其中所蕴含的情感和意图。近年来,情感分析在各个领域都得到了广泛应用,并取得了丰硕的研究成果。

1.1 情感分析研究

目前,情感分析研究方法主要分为 3 种,分别为基于情感词典的情感分析方法、基于传统机器学习的情感分析方法和基于深度学习的情感分析方法[5]。

基于情感词典的方法是依据情感词典中情感词的倾向来对语句的情感进行划分[6]。文献[7]在原有情感词典的基础上对 B 站领域的情感词典进行汇总和整理,运用 SO-PMI 算法进行情感倾向判断,实现面向 B 站领域情感倾向的分类系统。文献[8]采用 TF-IDF 训练结果和人工摄影情感基础词对领域词的权重进行修正,与 HowNet 情感词典合并形成摄影领域情感词典。

基于机器学习的情感分析方法是将文本数据转化为数字,在进行特征提取后对情感进行分析。文献[9]用层次支持向量机对微博文本进行多级情感分类。文献[2]用机器学习模型对外汇新闻进行分类,并建立程度词典细粒度地划分情感强度。文献[10]提出一种基于双语词典的多类情感分析模型,有效解决了现有情感词典多基于单一语言的问题。

在基于深度学习的情感分析方法中,文献[11]构建 BERT-CNN 模型和 BERT-RCNN 模型对微博中网友的情感进行识别并对比分析结果,文献[12]将 LDA 主题向量嵌入 BERT 词向量模型中,在 BERT 预训练模型的基础上添加 CNN,以准确地进行舆论情感分类。

本文采用基于词袋模型和 Word2vec 模型的决策树、朴素贝叶斯和支持向量机模型,并对它们的性能进行对比评估。

1.2 情感强度研究

文献[13]在程度词典的基础上结合外汇新闻领域特征词,计算外汇新闻相关语句的情感强度,计算强度与预期结果相近,但是该研究仅针对外汇新闻,对政务方面的新闻适应性有待提高。文献[3]以商品评论为研究对象,构建基于 PAD 模型的网络口碑情感强度测度模型,该模型对商品评论有较好的评测效果。文献[4]提出商品评论分析策略来分析情感倾向,并构建领域词典,将文本的情感强度进行量化。但是,以上 2 个研究的局限在于所构建的情感词典中词汇样本不够丰富,对网络词汇识别不准确。文献[14]提出基于情感本体的在线评论情感极性及强度分析方法,实现了在线评论的评论整体和具体特征的情感极性及强度分析,但其观点词情感的计算结果依赖于网络评论,没有与语言学专家标注的情感词典进行有效结合。

现有研究中所提供的数据对于政务新闻的适应性并不理想,且使用的文本词汇较为有限,难以准确识别网络用语。在进行情感结果计算时,未能与专家的情感词典进行有效结合,也影响了结果的准确性。

针对以上工作的不足,本文使用文献[15]整理和标注的大连理工大学中文情感词汇文本,并在此

基础上扩展政务微博领域的相关词汇。此举有效解决了先前研究在适应政务新闻、准确识别网络词汇以及与语言学专家情感词典相结合等方面的问题。

2 政务微博情感分析的影响因素及解决方案

2.1 政务微博情感分析的影响因素

政务微博情感分析的结果受到文本内容和语言风格的影响[16]。其中，数据质量问题和情感强度量化问题是影响分析结果准确性和可靠性的 2 个主要因素。

在划分网络热词、处理无关词汇和表情包时，数据的来源、采集方法和清洗方式等因素都会对分析结果产生直接或间接的影响。在计算情感强度时无法量化语言的情感强度，也会对分析结果产生影响。因此，为了提高政务微博情感分析的准确性和可靠性，需要充分考虑数据质量问题和情感强度量化问题，并采取相应的措施来解决这些问题。

2.1.1 数据质量问题

在政务微博情感分析中，网络热词的精准划分对于文本整体意思的理解具有至关重要的作用，若未对其进行准确划分，则会影响模型对文本情感强度和情感倾向的判断，从而影响整体的准确性。相比之下，无关词汇和表情包对于文本的整体意思并没有实质性的贡献。然而，通过爬虫获取的数据可能存在乱码问题，这会给情感倾向分析和强度分析带来阻碍，进而影响模型的准确性。因此，在政务微博情感分析中，需要充分考虑这些因素，并采取相应的措施来提高分析结果的准确性和可靠性。

2.1.2 情感强度量化问题

情感强度无法量化这一问题将导致得出的非量化指标无法准确判断情感的强弱程度，从而给工作人员的操作带来困难，难以直观地评估情感分析的结果，这将不可避免地导致在实际工作中出现误差，影响分析结果的准确性和可靠性。因此，在政务微博情感分析中，需要充分考虑这一问题，并探索有效的解决方案来提高分析结果的精度和可操作性。

2.2 影响因素的解决方案

针对数据质量问题，需要去除一些噪声和干扰因素，使得文本更加纯净和准确，从而提高情感分析的准确度，为此，本文采用了多种方法进行研究。首先，在预处理阶段，本文爬取与冬奥会、国足有关的数据，侧重考虑这两方面的微博热词、网络用语，在构建语料库时，考虑到爬取的数据与冬奥会、国足有关，因此侧重考虑这两方面的微博热词、网络用语。本文涵盖了冬奥会的运动项目和国足的一些常用词、相关运动员名字以及一些新兴的网络词语等，以确保所收集的热词尽可能完整和准确。之后，建立停用词表，并在数据清洗时使用该停用词表将文中的停用词进行剔除。接着，使用正则表达式将非中文文本进行去除，通过设计合适的表达式减少表情包对文本分析的影响。

为了提高情感分析的准确性和可靠性，本文结合了大连理工大学中文情感词汇文本，设计了一个带有基本权重值的情感强度计算公式，其中利用标签和文中词语的权重计算情感强度，以数字的绝对值大小进行表示。绝对值越大表示情感强度越高，反之则表示情感强度越低。这些情感词汇经过专家的整理和标注，具有较高的准确性和可靠性。使用这些情感词汇可以更准确地反映文本的情感色彩，提高情感分析的精度和可靠性。

3 政务微博情感倾向与情感强度模型

本文提出的基于支持向量机的微博评论文本情感倾向判断和情感强度计算框架，旨在预测微博文本的情感倾向，并提取影响情感强度的特征，结合强度词典分析评论的情感强度。

本文框架的实现主要分为以下步骤：使用爬虫获取微博文本；清洗获取到的数据；中文分词和词典构建；去停用词；文本向量化；利用支持向量机判断情感倾向；结合情感强度词典对文本的情感强度进行判断。通过该模型可以得出相应的情感数值，其中正值表示正向情感，负值表示负向情感，绝对值的大小表示情感强度的大小，绝对值越大表示情感越强烈，反之则情感越平淡。情感分析模型框架如图 1 所示(彩色效果见《计算机工程》官网 HTML 版，下同)。

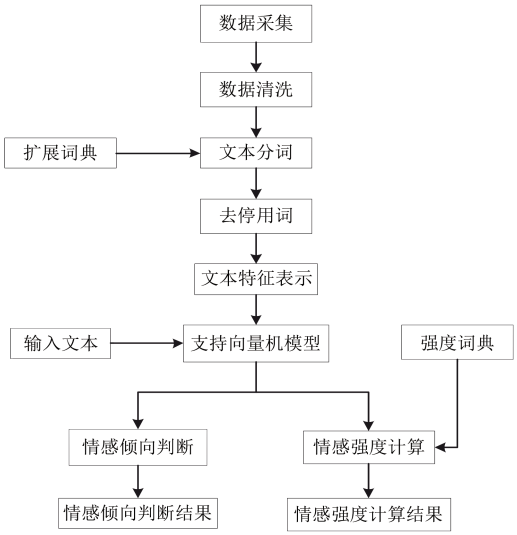

图 1 情感分析模型框架

Fig. 1 Framework of sentiment analysis model

3.1 数据采集

为了获取政务微博中的数据[17]，本文通过 Python 语言爬取新浪微博中 2022 年北京冬奥会和中国足协 2 个微博上网友的评论，这些评论被选为研究对象，旨在探究政务微博中公众情感表达的特点和规律。通过这种数据收集方式，能够获得更丰富、更具有代表性的数据，从而更好地进行情感分析和情感计算研究。文中爬虫实现思路如图 2 所示。

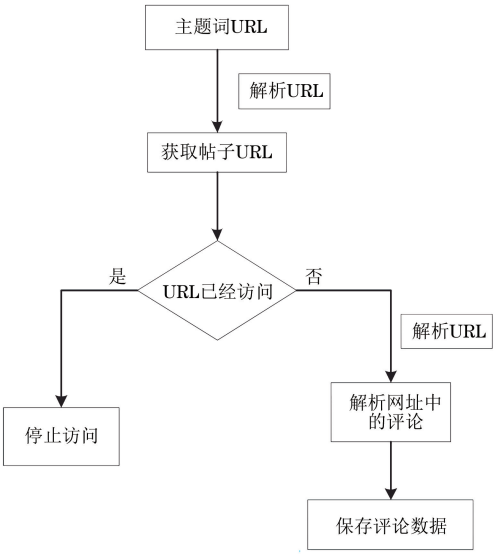

图 2 爬虫实现思路

Fig. 2 Implementation ideas for Web crawlers

为了解决评论内容分散在不同页面的问题，本文对爬虫步骤进行改进：确定主题，即冬奥会/中国男足；获取主题词的 URL 后进行解析，从而得到帖子的 URL；判断该帖子的 URL 是否已经访问，如果已经访问则停止，反之则解析帖子 URL，获取评论内容，保存评论内容。

本文在爬虫过程中遇到来自微博的反爬机制，通过研究发现是由于访问频率过高、访问 IP 过于固定等原因造成。为了规避这些限制，本文将访问频率控制在微博平台可以接受的范围内，即 1 次/min。此外，针对微博平台封锁 IP 的问题，本文采用了 IP 代理技术。最终，本文共爬取 29 331 条评论作为研究数据。

3.2 数据清洗

本文所获取的评论数据包含空格、换行符、表情包等众多非标准式的特殊字符，这些字符对于整体的评论语义并没有太大作用，因此，需要对特殊字符数据进行清洗和删除，只保留中文文本，从而使得文本更加规范化[18]，此举有助于提高数据质量。

为保证文本中只存在中文文本，使用正则表达式进行数据清洗。Unicode 为各种语言的字符设定了统一且唯一的二进制编码，能够满足跨平台、跨语言的文本处理和转换需求，并且汉字在 Unicode 编码中为一段连续的编码。因此，使用一段十六进制编码的正则表达式[^\u4e00-^\u9fa5]对数据进行清洗[19]。数据清洗前后效果对比如图 3、图 4 所示。

图 3 未进行清洗的数据(部分)

Fig. 3 Unclean data (partial)

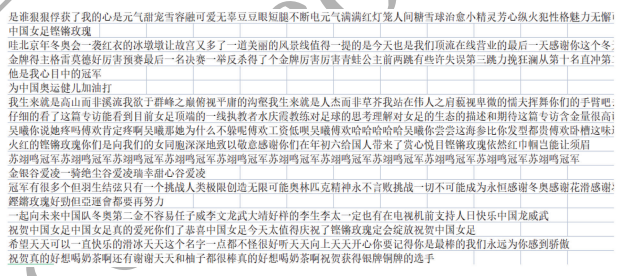

图 4 进行清洗后的数据(部分)

Fig. 4 Data after cleaning (partial)

3.3 中文分词与词典构建

从语言结构的角度来看，字构成词，词组成句。英文文本中的单词之间由空格作为分隔符来隔开，可以直接进行分词。但汉字组成的词语之间没有明确的分隔符，因此在处理中文文本时，精确地进行分词是基本且重要的工作。

在进行分词时，通常会直接使用基于统计分词算法的 Jieba 分词工具进行分词[20]。但是，Jieba 分词默认使用的分词词典可能无法准确识别政务领域相关的评论以及一些新兴词汇。因此，本文构建了与微博热词相关的语料库，通过此库再使用 Jieba 分词工具的精确模式进行分词。在这种模式下，分词工具可以更精确地对评论数据进行分词处理，从而尽可能地保持网络热词的完整性。

基于情感词典的情感分类效果取决于词典构建的完善程度，基于机器学习的情感分类效果取决于标注标签的正确性[21]。政务微博中体育相关公众号的话题评论能比较清晰合理地分辨情感倾向，因此，本文通过机器学习模型进行训练并预测政务微博评论的情感倾向，在预测情感倾向的基础上，结合程度词词典和情感词词典来计算政务微博评论的情感强度。

微博评论中的网络热词有“笑死”、“墩墩”、“好

牛”等。扩充语料前后进行分词时的区别如图 5 所示。

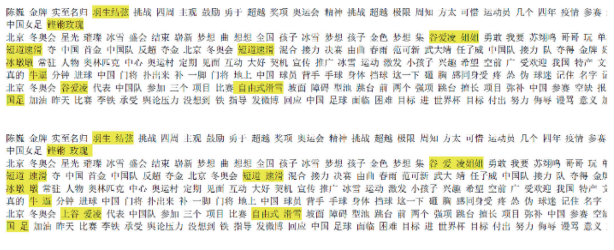

图 5 扩充语料前后进行分词时的区别

Fig. 5 The difference in word segmentation before and after expanding the corpus

3.4 去停用词

去停用词指的是在信息检索中，为了节省储存空间和提高搜索效率，在处理文本数据之前将某些没有实意且不能单独表达文本数据中信息的词过滤掉，此类词有介词、连接词以及一些副词等。通过人工手动输入停用词并收集整理为停用词表。本文所用的停用词有“有时”、“果然”、“某个”、“每天”、“比如”等，在进行数据分词的过程中，需要参考所建的停用词表。

3.5 文本特征表示

机器学习模型无法使用文本数据进行训练，模型只能够接收具有一定格式的、能够表达相应语义信息的数值化向量。因此，本文需要将文本数据转换成数值向量，形成一定结构的特征向量。词为中文文本语言的基本单位，正确地表示特征词对于模型训练结果至关重要。本文采用词袋模型和 Word2vec 模型对特征进行表示。

3.5.1 词袋模型

词袋模型是一种常用的表达文本特征的数学模型，该模型在 one-hot 编码的基础上对整个文本中词出现的频次进行记录，通过词语出现的频次来表示该词语对文本的重要程度。

在文本中有很多没有意义但出现频率很高的词语，此类词语会获得较大的权重。为避免停用词带来的权重影响，使用 TF-IDF 方法对词频进行加权处理[22]，这在一定程度上为词语的重要程度提供了合理性度量[23]。

$$V_{\mathrm{TF}}=\frac{M(w,S)}{M(S)} \tag{1}$$

$$V_{\mathrm{IDF}}=\log_a\left(\frac{P}{P(w)}\right) \tag{2}$$

$$V_{\text{TF-IDF}}=V_{\mathrm{TF}}\cdot V_{\mathrm{IDF}} \tag{3}$$

式中：V_{TF} 表示词频，即词语 w 在文档 S 中出现的次数，这个比例越高，说明词语在该文档中越重要；V_{IDF} 是逆文档频率；P 是语料库中文档的总数；$P(w)$是包含词语 w 的文档数目，对数函数用于降低频率的影响，使得 V_{IDF} 值不会增长得太快，V_{IDF} 的值越高，说明词语 w 在语料库中越稀有，因此越重要；$V_{\text{TF-IDF}}$ 是 V_{TF} 和 V_{IDF} 的乘积，它结合了词频和逆文档频率 2 个因素，用以评估词语在文档中的重要性，一个词语的 $V_{\text{TF-IDF}}$ 值越高，说明它在文档中越重要，同时在整个语料库中出现得越少，因此可以作为区分文档主题的一个有效特征。

在计算 TF-IDF 值后，按照值大小进行降序处理，选取前 500 个作为特征词，用于构建维度为 500 维的文本向量。

3.5.2 Word2vec

Word2vec 可以根据给定的文本数据，通过训练模型快速地将一个词语表达成向量形式，从而表达该词的更深层次特征，并且达到聚类分析、预测词之间相似性的目的[24]。

Word2vec 的网络结构与神经网络模型类似，包括输入映射层、隐藏层和输出层 3 层网络结构，其使用 one-hot 编码作为输入，通过神经网络的非线性结构映射出词的深层次语义信息，表达成低维向量。

Word2vec 包括连续词袋模型(CBOW)和 Skip-gram 这 2 种模型训练方法。连续词袋模型的思想是用中心词前后的 c 个词来计算这个中心词出现的概率。Skip-gram 模型与 CBOW 模型正好相反，通过输入中心词，利用神经网络计算出它的上下文。本文采用 CBOW 训练方法生成维度为 300 维的词向量。

3.6 支持向量机

支持向量机是一种广泛应用于分类、回归和离群点检测等领域的机器学习算法[25]。本文模型使用了支持向量机进行情感倾向判断以及情感强度计算，具体而言，使用支持向量机来构建一个分类器，用于对政务微博的情感倾向判断和情感强度计算。

在情感倾向判断方面，使用支持向量机训练一个分类器，每个类别对应一种情感倾向。训练数据包括政务微博的内容和标签(情感倾向)。通过训练，分类器可以学习到不同情感倾向的文本之间的差异，并根据输入的文本内容进行情感倾向预测。

在情感强度计算方面，使用支持向量机训练的分类器进行政务微博的情感强度计算，可以根据输入的文本内容预测其情感强度。

将支持向量机用于对政务微博的情感倾向判断和情感强度计算时，可以提供更准确的情感分析结果，为相关政策措施的制定提供有力的数据支持。

3.6.1 基于支持向量机的情感倾向判断

在微博等社交媒体平台上，网民们对各种话题的评论能够真实地反映自身情感[26]。判断情感倾向的方法有多种，比如情感词典模型、传统机器学习模型、深度学习模型等。本文对比了 3 种机器学习模型(支持向量机、朴素贝叶斯、决策树)，引入 2 种词向量(TF-IDF 加权后的词袋模型、Word2vec 词向量模型)，在扩充语料库前和扩充语料库后这 2 种情况下，模型准确率结果如表 1、表 2 所示。可以看出，支持向量机的模型准确率普遍较高，所以本文采用支持向量机用于情感倾向的判断，其目的在于预测输入句子的标签，预测结果分为正向和负向两类，分别使用 1 和 0 表示。

表 1 扩充语料库之前模型准确率结果

Table 1 Results of model accuracy before expanding the corpus

指标	支持向量机		朴素贝叶斯		决策树	
	TF-IDF	Word2vec	TF-IDF	Word2vec	TF-IDF	Word2vec
acc	0.825 4	0.843 1	0.820 4	0.722 9	0.809 1	0.716 4

表 2 扩充语料库之后模型准确率结果

Table 2 Results of model accuracy after expanding the corpus

指标	支持向量机		朴素贝叶斯		决策树	
	TF-IDF	Word2vec	TF-IDF	Word2vec	TF-IDF	Word2vec
acc	0.827 1	0.843 0	0.822 5	0.742 8	0.811 2	0.708 2

根据表 1 和表 2 的数据可以得出：在机器学习模型的准确率评测中，扩充语料库能够提高这些模型的准确率。具体而言，扩充语料库后，支持向量机的表现最为出色，准确率达到了 0.827 1，决策树的表现相对较差，仅为 0.708 2。这表明通过扩充语料库，可以提供更多的数据样本，有助于机器学习模型更好地学习文本特征和规律，从而提高准确率。因此，对于其他类似的任务，也可以考虑通过扩充语料库来提高模型的准确率。

3.6.2 基于支持向量机的情感强度计算

政务微博文本中存在大量带有情感的句子，本研究提出了一种模型，用于对文本进行评估并返回该文本的标签。该模型利用标签和文中词语的权重计算情感强度，以数字的绝对值大小来表示，绝对值越大表示情感强度越高，反之则表示情感强度越低。

1)情感强度。

情感强度是指中文文本情感倾向的强度。本文将用来分析情感强度的词语分成两类：第一类是程度词，表示一个词语的极性，如“超”、“很”、“较”、“一般”、“一些”等；第二类是普通词，该词本身也带有一定的情感强度同时可以被程度词修饰，这类词语也会根据本身情感强度赋予权值[27]。第二类词语数量很大，不易列举和收集，本文在大连理工大学中文情感词汇文本的基础上对政务微博领域的词汇进行了扩展。

2)强度计算。

考虑到程度词不会存在于每一条文本中，因此为其设置一个基本权重值，计算公式如下：

$$V=\begin{cases}W_{\text{程度词}}\cdot W_{\text{普通词}}, & \text{有程度词修饰}\\ W_{\text{基本权重}}\cdot W_{\text{普通词}}, & \text{没有程度词修饰}\end{cases} \tag{4}$$

式中：$W_{\text{程度词}}$ 表示程度词的权值，本文将该类词语分为 DV1、DV2、DV3、DV4 这 4 大类，并分别赋予权值 2、1.75、1.5、0.5，如表 3 所示；$W_{\text{基本权重}}$ 表示普通词没有程度词修饰时的权重，本文将其权值设置为 1；$W_{\text{普通词}}$ 表示普通词的权重，在扩展大连理工大学中文情感词汇文本中，也将普通词的权重划分为 5 类，其权值分别为 1、3、5、7、9，如表 4 所示。

表 3 程度词权重

Table 3 Weight of degree words

程度词	类别	权值
非常、极大、完全等	DV1	2
很、格外、相当等	DV2	1.75
还、过于、较等	DV3	1.5
稍微、大约、恰恰等	DV4	0.5

表 4 普通词权重

Table 4 Weight of common words

普通词	权值
灯火辉煌、不武断、分文不取等	1
信守、仙女、扒拉等	3
龙翔凤舞、铿锵玫瑰、群英汇等	5
宏才大略、下头、神仙等	7
景仰、无所不可、荣耀等	9

整个句子的情感强度计算如下：

$$S=(V_1+V_2+\cdots+V_n)\cdot P_i \tag{5}$$

式中：P_i 表示句子 i 的情感倾向，若句子 i 表示负面情感，则 P_i 值为 -1，若句子 i 表示正面情感，则 P_i 值为 1，该值取决于训练好的机器学习模型对句子标签的判断输出；V 表示句子中一个带有情感倾向的普通词的情感强度；S 为 V 的累加，表示整个句子的情感强度。

4 实验结果及分析

4.1 实验数据集

本文通过网络爬虫技术获取了 2022 年北京冬

奥会和中国足协 2 个微博公众号评论共 29 331 条，在 Excel 中对每条数据进行标记，将正向情感标记为 1，负向情感标记为 0。接着对数据进行排序操作，按照标签将评论分为正向和负向两类，然后对正向和负向评论进行数据清洗操作。结果显示，正向情感倾向的数据共有 19 993 条，负向情感倾向的句子总有 9 338 条。接下来，在 Excel 中计算 2 类文本每条评论的长度，再将 2 类文本按照评论长度进行降序处理。

由于微博中的评论大多为短文本数据，并且长短参差不齐，有的文本只有两三个字，如“嗯嗯”、“哈哈”、“呵呵呵”等，这些短文本词语并不能清晰地表达出情感倾向；有的文本长达上百字，但这些长文本在整个数据集中占比很小，不能很好地表示微博文本的特征。为避免数据集长度对分类效果产生影响，本文对正向文本数据和负向文本数据的长度进行筛选，去除一些短文本数据和长文本数据。该数据集所有数据的长度均在[0,140]这个区间内，采用 5 为步长对该区间进行划分。通过 Excel 统计长度为 5 的区间间隔内的文本频数，生成正向文本和负向文本的频率-长度直方图，分别如图 6、图 7 所示。

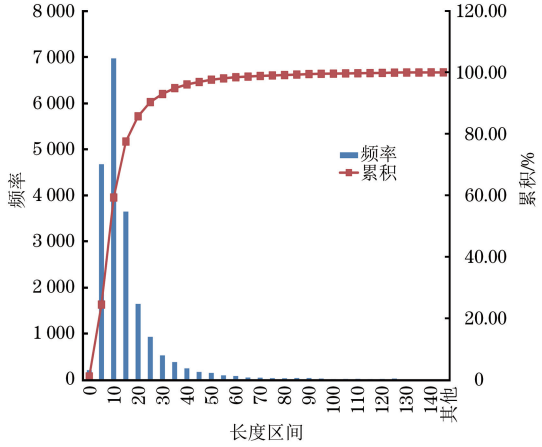

图 6 正向文本的频率-长度直方图

Fig.6 Frequency-length histogram of forward text

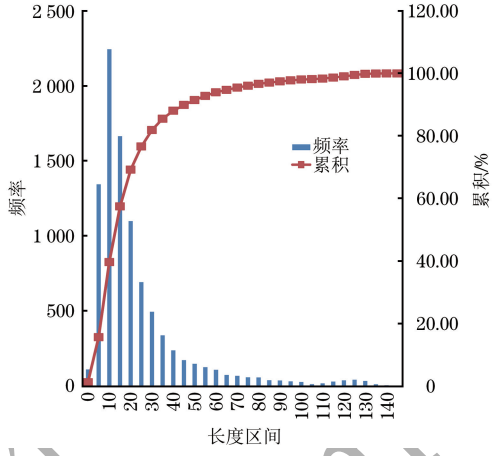

图 7 负向文本的频率-长度直方图

Fig.7 Frequency-length histogram of negative text

对出现频率高的长度区间内的数据进行筛选和保存，本文选择的长度区间为[5,100]。在对文本进行初步筛选后，正向文本有 16 508 条数据，负向文本有 8 172 条数据，分别进行保存。本文对筛选后的文本进行中文分词、去停用词操作后，将文本数据进行特征表示（词袋模型、Word2vec）：在使用词袋模型表示特征时，该特征向量的维度为 500 维，并且使用 TF-IDF 计算向量的权重；在使用 Word2vec 表示特征时，特征向量维度为 300 维。

4.2 情感分类评测方法及结果分析

4.2.1 情感分类的评测方法

本文选择了查准率（Pre）、查全率（Rec）和 F1 值这 3 个指标对情感倾向进行测评。为进一步展示分类器效果，本文使用 Python 函数生成混淆矩阵视图，以观察各类的分类情况。混淆矩阵中每行之和表示该类的样本数量，每列的值则代表被分到该类的数量。

4.2.2 情感分类的结果分析

本节对比分析了 3 种机器学习模型（支持向量机、朴素贝叶斯、决策树）引入 2 种词向量（TF-IDF 加权后的词袋模型、Word2vec 词向量模型）后，在扩充语料库前和扩充语料库后这 2 种情况下的实验结果，如表 5、表 6 所示。

表 5 扩充语料库之前模型评测结果

Table 5 Model evaluation results before expanding the corpus

类别	评估指标	支持向量机		朴素贝叶斯		决策树	
		TF-IDF	Word2vec	TF-IDF	Word2vec	TF-IDF	Word2vec
1	Pre	0.832 3	0.857 0	0.840 1	0.810 1	0.827 7	0.784 7
	Rec	0.931 8	0.923 9	0.909 8	0.775 7	0.907 9	0.803 4
	F1	0.879 2	0.889 2	0.873 5	0.792 5	0.866 0	0.793 9
0	Pre	0.803 2	0.803 8	0.764 5	0.558 8	0.750 6	0.555 4
	Rec	0.597 0	0.669 2	0.628 5	0.609 7	0.594 4	0.526 9
	F1	0.684 9	0.730 4	0.689 9	0.583 2	0.663 4	0.540 8

表 6 扩充语料库之后模型评测结果

Table 6 Model evaluation results after expanding the corpus

类别	评估指标	支持向量机		朴素贝叶斯		决策树	
		TF-IDF	Word2vec	TF-IDF	Word2vec	TF-IDF	Word2vec
1	Pre	0.833 4	0.856 6	0.841 8	0.805 0	0.831 4	0.780 2
	Rec	0.932 0	0.924 5	0.911 0	0.822 6	0.907 0	0.797 5
	F1	0.880 2	0.889 3	0.875 0	0.813 7	0.867 5	0.788 8
0	Pre	0.804 9	0.804 7	0.768 1	0.600 6	0.751 9	0.543 7
	Rec	0.601 8	0.668 0	0.632 5	0.572 5	0.605 3	0.517 7
	F1	0.688 7	0.730 0	0.693 7	0.586 2	0.670 7	0.530 4

通过对比表 5 与表 6 可以得出，在扩充语料库之后，3 种机器学习模型无论在 TF-IDF 加权后的词袋模型还是 Word2vec 词向量模型下，均表现出一定的提升。特别是在 Word2vec 词向量模型下，朴素贝叶斯机器学习模型中类别 1 的召回率和类别 0 的准确率提升明显，两者分别提升 0.046 9 和 0.041 8，但也是在此类模型的训练下，出现了一定程度的下降，类别 0 的召回率下降了 0.037 2。在 Word2vec 词向量模型下的决策树中，在扩充语料库之后各个数据均出现了下降。

由图 8 中的 6 个混淆矩阵对比可以看出，支持向量机在 TF-IDF 加权后的词袋模型作为特征的情况下，正确预测出 0 标签的个数比 Word2vec 作为特征多了100个，正确预测出1标签的个数比 Word2vec 作为特征少了 100 个。在 TF-IDF 加权后的词袋模型作为特征的情况下，支持向量机比决策树、朴素贝叶斯预测正确的个数分别多 100 个和 200 个。对比 Word2vec 模型下的 3 种模型，也能得出类似的结果。

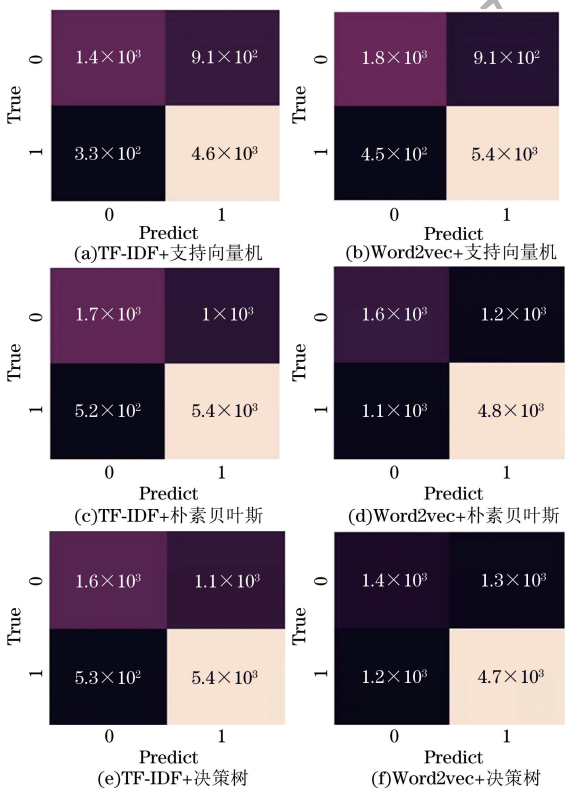

图 8 混淆矩阵对比视图

Fig.8 Comparison view of confusion matrix

综合对比支持向量机、朴素贝叶斯和决策树 3 种模型的性能，可见支持向量机的效果最优。此外，支持向量机在 TF-IDF 与 Word2vec 下的表现也很接近。经过以上讨论可以得出，本文提出的情感分析方法是可行的，即将词袋模型作为特征提取器并使用支持向量机模型进行预测具有有效性。

4.3 情感强度结果分析

使用本文提出的模型对政务微博中体育相关文本进行文本标签预测，预测标签如果为 1，则情感强度值为正；预测标签如果为 0，则情感强度值为负。实验结果如表 7 所示。

表 7 情感强度计算结果

Table 7 Results of emotional intensity calculation

句子	情感强度
冰墩墩很可爱，给我 rua 两下	8.75
恭喜中国队真的超级棒超级辛苦了	10
没有一种成功的背后没有努力做支撑	3
麻烦给小苏申诉在家门口举办还被明着欺负有没有一点骨气啊	−15
自己家门口的比赛，自家孩子被打压，你们在干嘛？别成天整那些和谐友善，先把自家运动员的应有权利争取到吧！	−10

百度智能云可以输出消极概率、积极概率，两者相加之和为 100%，由此可以将概率与置信度相结合，并将计算得到的结果映射到[−15，15]，从而计算出文本的情感强度，并与自建模型进行比较。

通过对比表 7 与表 8 可以发现，情感强度所反映的情感倾向与表 5、表 6 中的情感倾向具有一致性，这表明 2 个模型在其内部具有一致性。针对表 3～表 5 显示的百度智能云所计算出的情感强度

较为“极端”的情况，其表现在数值分布于映射集合的两端，这很大可能产生忽略中性文本的问题。相反地，本文所提模型能够更好地分散数据，从而解决忽略中性文本的问题，证明了本文所提模型的优越性和应用潜力。因此，可以认为本文模型在情感分析领域具有显著优势。

表 8 百度智能云的情感强度计算结果

Table 8 Calculation results of emotional intensity of Baidu AI cloud

句子	情感强度
冰墩墩很可爱，给我 rua 两下	14.23
恭喜中国队真的超级棒超级辛苦了	14.99
没有一种成功的背后没有努力做支撑	−14.67
麻烦给小苏申诉在家门口举办还被明着欺负有没有一点骨气啊	14.11
自己家门口的比赛，自家孩子被打压，你们在干嘛？别成天整那些和谐友善，先把自家运动员的应有权利争取到吧！	−14.27

5 结束语

本文通过对政务微博进行研究设计了一个关于舆情情感倾向和情感强度的模型，该模型解决了现有研究在政务新闻适应性、网络词汇识别准确性和与情感词词典相结合等方面的不足。通过爬取微博评论数据，并采用词袋模型和 Word2vec 两种文本特征表示方法，结合朴素贝叶斯、决策树和支持向量机 3 种机器学习模型进行情感倾向预测，验证了模型的可行性。在模型评估方面，通过混淆矩阵对比视图分析了不同模型的预测效果，结果表明，在 Word2vec 词向量模型下，支持向量机机器学习模型表现显著，验证了本文模型的有效性。此外，对于情感强度计算，本文将百度云情感分析的置信度进行了映射，并与本文模型计算的情感强度进行对比，发现本文模型计算出的情感强度能够更准确地反映中性文本的情感属性。最后，应用所提模型对政务微博文本进行情感倾向和情感强度的计算，结果表明，模型的计算结果符合人们对相关文本的情感倾向和情感强度感知，可以为政府在网络舆论的筛选排查和管控方面提供参考。下一步将对多源数据融合、实时监测与预警系统开发等方面进行研究，旨在进一步提高模型的准确性和实用性。

参考文献

[1] 中国互联网络信息中心. 中国互联网络发展状况统计报告[J]. 国家图书馆学刊，2023，32(2)：1-39.
China Internet Network Information Center. Statistical report on the development of Internet in China[J]. Journal of the National Library，2019，32(2)：1-39. (in Chinese)

[2] 戚天梅，过弋，王吉祥，等. 基于机器学习的外汇新闻情感分析[J]. 计算机工程与设计，2020，41(6)：1742-1748.
QI T M，GUO Y，WANG J X，et al. Sentiment analysis of foreign exchange news based on machine learning[J]. Computer Engineering and Design，2020，41(6)：1742-1748. (in Chinese)

[3] 李吉，黄微，郭苏琳，等. 网络口碑舆情情感强度测度模型研究——基于 PAD 三维情感模型[J]. 情报学报，2019，38(3)：277-285.
LI J，HUANG W，GUO S L，et al. Research on the sentiment intensity measurement model of Internet word-of-mouth public opinion based on the PAD model[J]. Journal of the China Society for Scientific and Technical Information，2019，38(3)：277-285. (in Chinese)

[4] 孔伟俊，胡广朋. 基于领域词典的网络商品评论情感分析[J]. 计算机与数字工程，2018，46(1)：155-159.
KONG W J，HU G P. Analysis of Internet product reviews which based on field of emotional dictionary[J]. Computer & Digital Engineering，2018，46(1)：155-159. (in Chinese)

[5] HU Y. Text mining and data information analysis for network public opinion[J]. Data Science Journal，2019，18：7.

[6] 吴杰胜，陆奎. 基于多部情感词典和规则集的中文微博情感分析研究[J]. 计算机应用与软件，2019，36(9)：93-99.
WU J S，LU K. Chinese Weibo sentiment analysis based on multiple sentiment lexicons and rule sets[J]. Computer Applications and Software，2019，36(9)：93-99. (in Chinese)

[7] 杨廉正，翟天智. 基于情感词典的视频评论情感倾向分析研究[J]. 网络安全技术与应用，2022(3)：53-56.
YANG L Z，ZHAI T Z. Analysis and research on emotional tendency of video comments based on emotional dictionary[J]. Network Security Technology & Application，2022(3)：53-56. (in Chinese)

[8] 刘亚桥，陆向艳，邓凯凯，等. 摄影领域评论情感词典构建方法[J]. 计算机工程与设计，2019，40(10)：3037-3042.
LIU Y Q，LU X Y，DENG K K，et al. Construction method of sentiment lexicon for photography reviews[J]. Computer Engineering and Design，2019，40(10)：3037-3042. (in Chinese)

[9] 李绍华，冯晶莹，张皓泓，等. 基于支持向量机的微博评论舆情分析[J]. 大学图书情报学刊，2021，39(5)：110-116.
LI S H，FENG J Y，ZHANG H H，et al. Public opinion analysis of microblog comments based on support vector machine[J]. Journal of Academic Library and Information Science，2021，39(5)：110-116. (in Chinese)

[10] 栗雨晴，礼欣，韩煦，等. 基于双语词典的微博多类情感分析方法[J]. 电子学报，2016，44(9)：2068-2073.
LI Y Q，LI X，HAN X，et al. A bilingual lexicon-based multi-class semantic orientation analysis for microblogs[J]. Acta Electronica Sinica，2016，44(9)：2068-2073. (in Chinese)

[11] 张苑，祝小兰，杨东晓. 基于深度学习的疫情情感分析[J]. 智能计算机与应用，2022，12(3)：40-45，52.
ZHANG Y，ZHU X L，YANG D X. Sentiment analysis of epidemic situation based on deep learning[J]. Intelligent Computer and Applications，2022，12(3)：40-45，52. (in Chinese)

[12] 辛明远，刘继山. 基于 BERTCNN-LDA 模型的舆情检测方法——以双减政策为例[J]. 信息与电脑(理论版)，2022，34(2)：59-63.
XIN M Y，LIU J S. Public opinion detection method based on BERTCNN-LDA model——a case study of double reduction policy[J]. Information and Computer(Theoretical Edition)，2022，34(2)：59-63. (in Chinese)

[13] 王吉祥，过弋，戚天梅，等. 嵌入互联网舆情强度的人民币汇率预测[J]. 计算机应用，2019，39(11)：3403-3408.

WANG J X, GUO Y, QI T M, et al. RMB exchange rate prediction based on Internet public opinion intensity[J]. Journal of Computer Applications, 2019, 39(11):3403-3408. (in Chinese)

[14] 郑丽娟，王洪伟. 基于情感本体的在线评论情感极性及强度分析：以手机为例[J]. 管理工程学报，2017，31(2)：47-54.
ZHENG L J, WANG H W. Sentimental polarity and strength of online cellphone reviews based on sentiment ontology[J]. Journal of Industrial Engineering and Engineering Management, 2017, 31(2): 47-54. (in Chinese)

[15] 赵鹏，何留进，孙凯，等. 基于情感计算的网络中文信息分析技术[J]. 计算机技术与发展，2010，20(11)：146-149，173.
ZHAO P, HE L J, SUN K, et al. Analyzing technologies of Internet Chinese information based on affective computing[J]. Computer Technology and Development, 2010, 20(11): 146-149, 173. (in Chinese)

[16] 李捷，袁周敏. 基于情感计算的政务微博情绪话语管理研究[J]. 外语教学，2023，44(5)：47-52.
LI J, YUAN Z M. Research on emotional discourse management of government micro-blog based on emotional computing[J]. Foreign Language Education, 2023, 44(5): 47-52. (in Chinese)

[17] RAJIV S, NAVANEETHAN C. An optimal topic centric crawler for acquiring bio-medical themes utilizing Gaussian support vector regression[J]. SN Computer Science, 2023, 4(6): 838.

[18] 靳宇倡，邓成龙，吴平，等. Emoji 图像符号的社交功能及应用[J]. 心理科学进展，2022，30(5)：1062-1077.
JIN Y C, DENG C L, WU P, et al. Emoji image symbol's social function and application[J]. Advances in Psychological Science, 2022, 30(5): 1062-1077. (in Chinese)

[19] BORSOTTI A, BREVEGLIERI L, CRESPI REGHIZZI S, et al. General parsing with regular expression matching[J]. Journal of Computer Languages, 2023, 74: 101176.

[20] 曾小芹. 基于 Python 的中文结巴分词技术实现[J]. 信息与电脑，2019，31(18)：38-39，42.
ZENG X Q. Technology implementation of Chinese Jieba segmentation based on Python[J]. China Computer & Communication, 2019, 31(18): 38-39, 42. (in Chinese)

[21] 万岩，杜振中. 融合情感词典和语义规则的微博评论细粒度情感分析[J]. 情报探索，2020，11(11)：34-41.
WAN Y, DU Z Z. Fine-grained sentiment analysis of microblog comments based on fusion of sentiment lexicon and semantic rules[J]. Information Research, 2020, 11(11): 34-41. (in Chinese)

[22] XIANG L. Application of an improved TF-IDF method in literary text classification[J]. Advances in Multimedia, 2022, 2022: 9285324.

[23] 杨欣，郭建彬. 基于改进 TF-IDF 的百度百科词语相似度计算[J]. 甘肃科学学报，2019，31(2)：143-147.
YANG X, GUO J B. Word similarity calculation of Baidu baike terms based on the improved TF-IDF[J]. Journal of Gansu Sciences, 2019, 31(2): 143-147. (in Chinese)

[24] MA Y Y, LIU C L, ZHANG J T, et al. Reliability study of stock index forecasting in volatile and trending cities using public sentiment—based on Word2vec and LSTM models[J]. Applied Economics, 2023, 55(43): 5013-5032.

[25] 王文韬，张士豹. 基于情感词典和 SVM 的微博网民情感分析[J]. 现代信息科技，2021，5(24)：24-27，31.
WANG W T, ZHANG S B. Emotion analysis of micro-blog netizens based on emotion dictionary and SVM[J]. Modern Information Technology, 2021, 5(24): 24-27, 31. (in Chinese)

[26] 王文静，郝其宏. 突发事件中政务微博网络舆情治理探究[J]. 国际公关，2023(21)：139-141.
WANG W J, HAO Q H. Research on the governance of online public opinion in government affairs microblog in emergencies[J]. International Public Relations, 2023(21): 139-141. (in Chinese)

[27] 李晶洁，胡奕阳，陶然. 基于情感倾向分析的语义韵强度算法探析[J]. 外国语(上海外国语大学学报)，2022，45(5)：65-74.
LI J J, HU Y Y, TAO R. Calculation of semantic prosody strength based on sentiment analysis[J]. Journal of Foreign Languages, 2022, 45(5): 65-74. (in Chinese)

编辑 吴云芳