DOI：10.13955/j.yzyj.2022.02.06.05

文本情感分析技术在中邮网院的应用研究

苗文凯，刘庆芳，刘海云，苏健

（石家庄邮电职业技术学院，河北　石家庄　050021）

摘　要：梳理了中邮网院在用的多个系统，以调查问卷系统为例，提出了文本情感分析技术在中邮网院的应用方案，并通过实验对方案进行了验证。

关键词：在线教育；文本情感分析；自然语言处理

中图分类号：F61　　**文献标识码：**A

中国邮政网络学院（以下简称“中邮网院”）支撑了邮政企业近百万员工的在线教育和培训，经过多年的积累和发展已经涵盖了邮务、寄递和金融三大板块的各类岗位和专业培训，在邮政人才培养过程中发挥着举足轻重的作用。随着邮政企业内部培训的不断深入和业务的扩展，与培训相关的数据越来越多，虽然中邮网院已经针对这些数据进行了全方位、多维度的统计分析，但是这些分析大多数都是基于结构化数据，对于论坛系统中的帖子和问卷系统中的建议等非结构化数据的利用有所欠缺。因此，挖掘这些非结构化数据内在的价值就显得尤为重要。

文本情感分析技术已经广泛应用于多个领域，它能对文本的情感倾向进行预测，进而为管理决策和个性化推荐等相关领域提供快速有效的分析结果。本文对中邮网院在用的多个系统进行梳理，使用情感分析技术处理非结构化数据，减少人工处理数据的工作量，快速挖掘非结构化数据中的价值，并依据情感预测的结果对相关业务进行优化，及时地调整培训内容和模式，提高教育培训的成效。

1　中邮网院相关系统介绍

中邮网院经过多年的发展已经形成了集在线培训和管理、资源管理和服务、人才测评和管理、继续教育和考试于一体的完整培训教育体系，主要包括 CMS 系统，邮政业务、邮储银行和中邮保险三大分院系统，直播系统，问卷系统，论坛系统，指标体系系统。在这些系统中都有数据分析和统计，其中指标体系系统是一个集大成的可视化分析统计系统，它把与培训相关的所有数据处理后进行多种类型（地图、饼图、折线图、时序图等）的可视化展示，让管理和业务人员对数据一目了然，帮助管理者熟知培训过程中的各类学习培训数据（参加

基金项目：邮政应用技术协同创新中心资助项目（项目名称：情感分析技术在中邮网院的研究和应用；项目编号：YB2021010）。

作者简介：苗文凯（1988 ～），男，河北石家庄人，硕士，工程师，主要从事个性化推荐、自然语言处理研究；刘庆芳（1981 ～），男，河北石家庄人，硕士，高级工程师，主要从事软件工程与远程教育研究；刘海云（1971 ～），男，北京人，硕士，研究员级高级工程师，主要从事信息技术、网络教育研究；苏健（1977 ～），男，河北泊头人，硕士，高级工程师，主要从事软件工程和数据库研究。

收稿日期：2021—11—04　　**本刊网址：**yzyj.sjzpc.edu.cn

培训班数量、学习时长、学习次数和考试情况等）。

2　文本情感分类技术相关研究

文本情感分析技术是指对互联网中海量的非结构化文本数据的感情倾向进行预测，预测结果可分为积极、消极和中性倾向，它属于自然语言处理的范畴，已经广泛应用于不同场景。情感分析技术主要实现方法有基于词典、基于机器学习和基于深度学习三类。基于深度学习的情感分析技术虽然能实现无监督的情感预测，但是训练神经网络模型需要大量的标注数据集并且耗时较长，因此本文采用基于词典和机器学习相融合的方法，这样即可在不丢失预测准确度的情况下，提高情感预测的效率。该方法处理的非结构化数据一般来源于互联网，通过网络爬虫来抓取指定的信息，经过数据预处理去除数据中的噪声和无关信息，进一步对数据进行分词和向量化，得到数据的特征向量后使用分类器对数据进行情感预测，主要流程如图 1 所示，主要包括数据分词和词性标注、特征值提取和情感预测等流程。

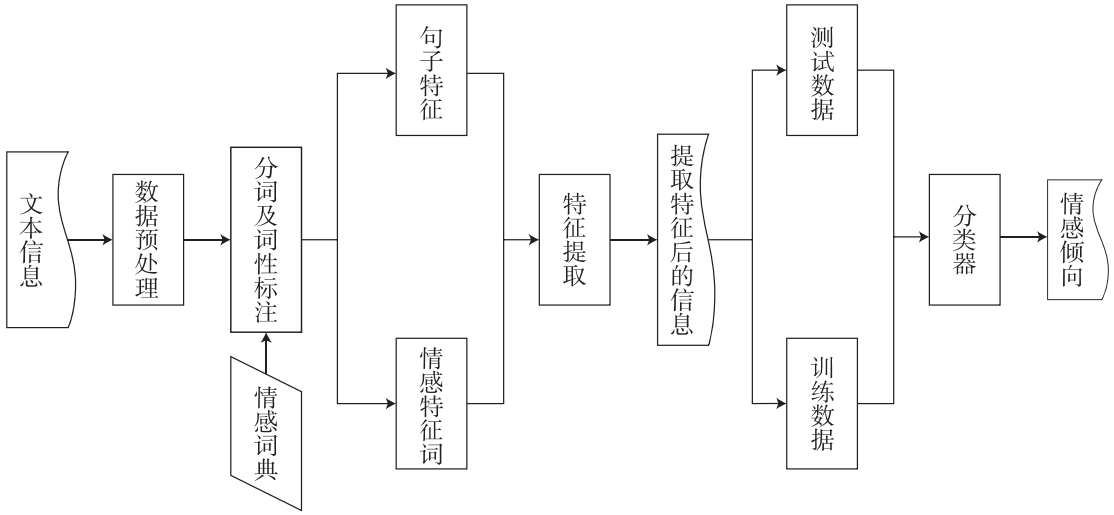

图 1　情感分析流程图

3　中邮网院调查问卷系统情感分析设计

中邮网院的调查问卷系统能够针对线上培训班开设调查问卷，也可以面向全学员开设调查问卷，以问卷的形式来收集学员对某一个培训班或者对某一类培训的满意度或者建议，通过对问卷结果的分析来提升培训质量，解决与学员工作息息相关的实际业务问题，为在线培训的主办方或者系统使用者提供重要改进方向和决策依据。问卷的内容主要包括客观的选择性项目和主观的意见建议，选择性项目只需学员在既定的选项中选择；对于主观的意见建议，学员可以根据自身的实际情况进行填写。本文在调查问卷的基础上，结合情感分析技术提出了一种业务培训模式，能够在较大程度上结合培训学员的实际情况进行培训内容的调整，最大程度地发挥培训的作用，提高培训的成效。该模式的主要流程如图 2 所示，在培训业务开展前后都进行问卷调查，通过对问卷非结构化数据的情感分析，可以获取问卷的情感值和情感文本，帮助管理者快速定位有价值的意见建议，为改进培训内容和培训方式提供依据。

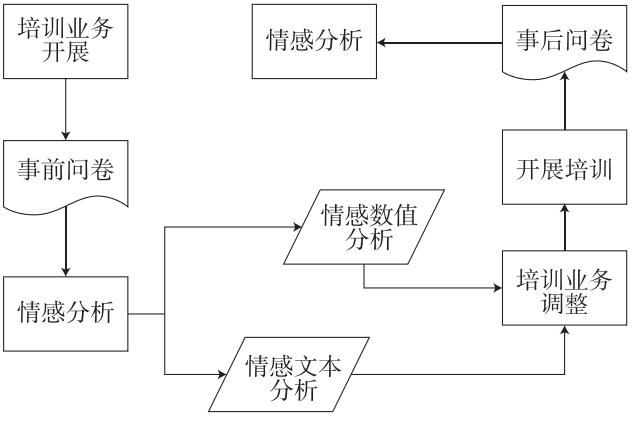

图 2　培训业务模式

3.1　问卷数据分词和词性标注

本文的数据来源是中邮网院问卷系统的数据库，通过数据库结构化查询语言，获取调查问卷的非结构化数据。首先使用正则表达式剔除乱码和无效数据，统一数据格式为 UTF-8 编码；然后使用分词工具对有效数据进行分词和词性标注（名词、动词、形容词、否定词等），在分词的过程中去除代词和副词等对情感预测几乎没有影响的词语，这样做可以减少后期特征向量的维度，降低算法的复杂度。

3.2　数据特征值提取

数据特征值的提取归根结底是要把非结构化数据转换成计算机能够识别的结构化、向量化数据，传统词袋（TFIDF 和词频等）模型在特征值提取的过程中会损失原数据的上下文信息，降低算法的准确度。尤其对于用来进行情感分析的数据，形容词在很大程度上能够改变句子的情感倾向，尤其是否定词，它甚至可以改变原有句子的情感色彩，所以数据的上下文信息对情感预测的结果影响较大。因此，本文选择使用 Word2vec 来实现信息的特征值提取，该算法能够保留词语的上下文信息，最大化保留原始信息，然后将信息转换成词向量的表达形式。比如句子“这次的培训班让我学到了很多，但是有些知识过于理论化，如果能够结合员工的实际工作业务，相信培训的效果会更好”，通过使用 Word2vec 计算后可以得到一个 N 维向量，该向量能最大程度代表非结构化数据的内容，做为后续情感分类的数据来源。

3.3　数据情感预测

情感预测本质上是一个分类问题，本文使用 SVM 算法来进行情感的分类预测，SVM 全称是 Support Vector Machine（支持向量机），它是一种基于统计学的机器学习算法，能够把数据表示在多维度的空间内，然后在这个多维度的空间内对数据进行分类。SVM 算法的核心是如何找到一个超平面把数据进行最大化的划分，一般情况下需要选择目标函数和约束条件，通过对大量数据的训练得到目标函数作为分类器，通过该分类器就可以对既定的数据进行分类。有研究表明，还可以给情感分析加上一个时间序列，随着时间变化针对某一个事件的情感倾向也会动态变化，这对持续追踪某一类事件来说具有重要意义。

4　实验和应用场景分析

4.1　实验环境和样本选择

硬件：Intel（R） Core（TM） i5-3470、内存 16G；软件：Python 3.6 和 PyCharm 集成环境，第三方工具主要使用 pandas、gensim 和 snownlp 等。本文采集了问卷系统中关于中邮网院的意见和建议的数据，经过数据预处理过程后，最终获得 11 000 多条有效的问卷数据，其中 70% 作为训练样本集，30% 用于验证集。

4.2　实验过程

pandas 是一款非常强大的数据处理工具包，该工具不仅能够读取、写入文件和矩阵运算，还能对数据进行可视化展示，本文使用它来方便快速地读取问卷数据，读取数据后使用 snownlp 工具进行分词和词性标注，得到详细的语料数据。snownlp 工具专门处理中文数据，其内置了很多训练好的数据集供用户使用，由于本文中涉及的问卷数据不包含特殊信息，所以使用内置的数据集能够在不丢失精度的情况下显著减少工作量。然后使用 gensim 工具包中的 Word2vec 算法对语料数据进行进一步处理，得到数据的特征向量，该特征向量能够被计算机识别进行数值运算，进而使用 SVM 分类算法对语料的情感倾向进行预测。情感分类是非线性的分类问题，所以在该过程中最重要的一步就是核函数的选择，经过多次的实验，本文选择使用高斯核函数，并且设置 C=20 和 Gamma=0.2 时的分类效果最好。

4.3　评价指标和结果

本文主要使用了精确度（P）、召回率（R）、F1-Measure 和准确率 Acc 四个指标进行评价，计算公式如下：

$$P=\frac{TP}{TP+FP} \tag{1}$$

$$R=\frac{TP}{TP+FN} \tag{2}$$

$$F1=\frac{2\times P\times R}{P+R} \tag{3}$$

$$Acc=\frac{TP+TN}{TP+FN+FP+TN} \tag{4}$$

其中，TP 表示样本为积极倾向的预测为积极；FN 表示样本为积极倾向的预测为消极；FP 表示样本为消极倾向的预测为积极；TN 表示样本为消极倾向的预测为消极。经过实验得出，该方法的精确度为 0.72，召回率为 0.87，准确率为 80.7%，F1 为 0.79，这个准确率基本符合生产系统中对于情感预测的标准。

4.4　实际应用场景分析

调查问卷系统中包含很多种类的问卷，有评价类型的问卷和意见建议类型的问卷，虽然设计者设计了众多的调查项目，但是一般情况下最后一项都是“其他”或者“请描述您的意见”这样的项目，以此来收集学员的主观描述，这类以文本形式存储的数据更能描述用户的真实需求。对于问卷中的客观选择类型的项目，问卷设计者可以明确地统计出项目的评分或者倾向，但是对于文本形式的内容，设计者无法在短时间内找到有价值的信息，只能通过人力进行主观判断，这就导致管理者需要花费大量的精力从大量的数据中寻找有价值的信息，工作量非常大。在以上的实验过程中发现：消极倾向的文本中往往包含更有价值的信息，更能明确地表达填写问卷者的真实想法。因此，文本的情感分析能够帮助问卷设计者在减少工作量的同时快速发现问卷中有价值的内容。

4.4.1　情感值分析

通过对问卷数据的情感分析得到每一条建议的情感值，情感值的范围为 0 到 1，0 代表消极倾向，1 代表积极倾向。通过对问卷数据的情感值统计（见图 3），可以发现学员对于中邮网院的整体情感倾向是积极的，0.5 以上的占据总数的 70% 左右，说明学员对中邮网院开展的各项教育培训的认可度较高。针对培训班的调查问卷也能得到情感值的统计，在培训班开展前进行一次问卷调查，调查问卷的主要内容是学员对该次培训的内容和形式的建议。培训班结束后再进行一次问卷调查，问卷内容主要是学员对培训的感想和培训的改进方向，通过对培训前后两次调查问卷进行情感分析，综合事前问卷和事后问卷的情感倾向，能够表现出学员对于该培训班的情感倾向，帮助管理者从宏观角度把握培训业务的方向，着重关注学员对培训的情感倾向，以提升培训的整体效能。

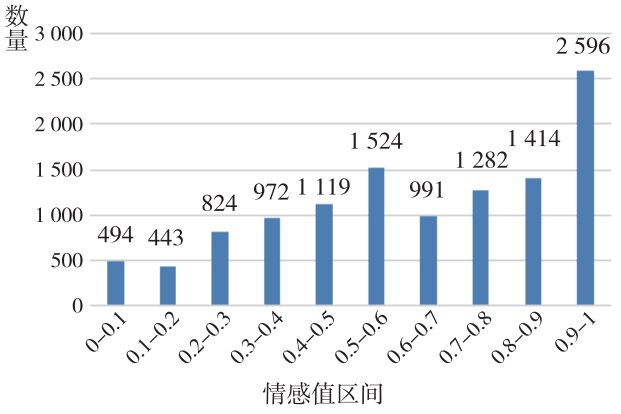

图 3　情感值区间统计

4.4.2　消极倾向数据分析

经过多次的实验得出结论，消极倾向的数据中潜在的价值更多，所以本文将情感值进行倒序排列，再依据二八定律，截取 20% 的数据进行分析，这 20% 的数据基本上概括了问卷的大部分内容，经过整理后可以快速得到部分消极倾向数据，如表 1 所示。

通过表 1 可以看出，情感值小的数据确实包含着学员在实际工作中遇到的问题和对培训内容的期待，这些意见和建议是培训过程中产生的宝贵财富，培训管理者要充分利用问卷数据，因地制宜根据员工的实际情况有针对性地开展培训教育工作，提升邮政企业员工的业务水平和综合素质，保障培训顺利开展。还可以把这些数据进行归类，包括技术类建议、业务类建议、管理类建议和政策类建议等，这些归类的数据经过长时间的积累将会是企业的过程资产，为企业培训的开展奠定坚实的基础，促进企业的可持续健康发展。

结语

本文使用情感分析方法对问卷数据进行实验分析，实验结果基本符合实际的生产情况。在实验研究过程中发现，消极倾向的文本中包含着更有价值的信息。情感分析技术很好地解决了管理者处理大量非结构化数据的痛点，能够快速为管理者找出数据中潜在的价值，为管理者和决策者提供有效的数据支撑，让管理者可以根据企业员工的实际情况组织和开展培训。在未来的工作中，可以结合学员的岗位、工龄等属性进行更深层次的数据挖掘和分

表 1　消极情感数据

序号	内容	情感值
1	邮政营业员基本没参加系统专业培训，农村网点业务量少，营业员不熟、不懂业务，建议多开设邮政营业新员工培训的课件。	0
2	课件时间不要太长，尤其是对邮政投递人员的培训，因为现在的邮件越来越多，投递时限越来越严格，投递人员没有多余的时间学习太长的课件。	0.000 03
3	好多投递员不熟悉业务，特别是快递包裹中转、交接，如快递包裹破损、水果腐烂、内件是液体且渗漏，没有统一操作流程。	0.000 03
4	其他快递运输的时限和邮件的破损情况，邮政邮件破损情况过多。	0.000 44
5	学习 110 监控中心或其他银行监控中心。	0.000 98
6	在顺丰平台，一些问题快件可以直接提交平台，如电话不接、手机空号、收件人恶意投诉，怎么处理？	0.001 07
7	用户下单地址与投递员投递范围不一致。	0.001 15
8	可以增加注册会计师、注册税务师、银行初级及中高级等相关培训。	0.001 4
9	在线答疑，在线客服。	0.001 82
10	在线问答可以有吗？有些疑难杂症的问题解决不了的时候，不知道问谁。	0.001 87
11	网上卖东西，可以增加一份寄递收入；同时在收寄过程中要先验视，再和寄件人眼同封装，实名收寄，实名登记，严格保密客户信息！	0.001 97
12	建议在投递环节提高投递员的快速性和安全性！在揽收环节提高合理性！	0.002 4
13	要实事求是，不能纸上谈兵，多考察实地情况。	0.002 55
14	建议中邮网院多点关于普邮投递、包裹投递等方面的内容，通过不断学习，提高投递员的业务水平。	0.002 6
15	针对邮政投递人员，应单独进行多次面对面的培训，提高邮政投递员的法律法规意识。现在邮政投递员的业务水平参差不齐。	0.002 76
16	提供学习赚积分，提供培训课件、视频下载渠道。	0.004 01
17	比如在新一代寄递平台上线时，只安排了根本不操作系统的网点负责人培训，而真正的系统操作生产经营人员，未提供任何培训。	0.004 16
18	学习积分制，证书颁发。纳入工资等级级别。	0.005 71
19	须经常给投递人员进行网上培训。	0.005 74
20	具有针对性的岗位培训资料。跳出邮政看邮政，跳出网院看网院，发现自己的不足，及时更改、补充、创新。	0.005 75

析；同时还需要完善情感词典，提高分类器的准确率，降低算法的复杂度，可以尝试使用深度学习相关的技术来实现情感的预测。

参 考 文 献

［1］王婷，杨文忠．文本情感分析方法研究综述［J］．计算机工程与应用，2021（12）

［2］张志．基于文本情感分析的快递企业物流服务质量评价研究［D］．合肥：合肥工业大学，2020

［3］王颖洁，朱久祺，汪祖民，等．自然语言处理在情感分析领域应用综述［J/OL］．计算机应用：1-12．http://kns.cnki.net/kcms/detail/51.1307.TP.20210928.1611.014.html，2021-12-22

［4］盛伟翔，扶齐彦．基于文本情感分析的大学生网络发帖调查［J］．电子技术与软件工程，2021（13）

［5］邓君，孙绍丹，王阮，等．基于 Word2Vec 和 SVM 的微博舆情情感演化分析［J］．情报理论与实践，2020（8）