NLP技术在文本情感分析中的应用研究

宋伟伟

（四川财经职业学院　四川　成都　610101）

[摘　要]文本情感分析在大数据时代发挥的作用越来越大，根据已有特定主题的态度、看法进行文本挖掘和分析，从而得到该评论或者看法是消极或者积极态度，有助于用户在繁杂的大量数据中做出比较合理的决策。在本文中，我们以电影的评论文本作为实验数据，通过特征选取、文本转换为特征变量、划分训练集与测试集、构建分类器、验证分类器等NLP技术对文本情感进行分析，计算出文本的情感态度从而为用户进一步决策提供依据。

[关键词]NLP技术；情感分析；分类；决策

【DOI】10.12252/j.issn.2096-627X.2020.02.1916

一、数据分析技术概述

（一）python语言

技术成熟的数据分析工具主要有Python、R、Stata、MATLAB、EViews、SAS、SPSS等。本文实验应用是使用python进行的。

Python是一种面向对象、解释性的计算机程序设计语言，python语言简洁而清晰，具有丰富和强大的类库[1]。在python中，通过安装Pandas、Numpy、Scipy、Statsmodels、Matplotlib、seaborn、Scikit-Learn、Theano、Tensorflow等一系列的程序包，本文实验中，通过Pandas实现数据分析、分组聚合等操作，Matplotlib实现数据可视化。丰富的类库满足绝大多数的应用需求。

（二）Anaconda开发环境

Anaconda是一个集成开发环境[2]，本实验是在Anaconda开发环境中进行，Anaconda是一个开源的python发行版本，包含了conda、python等180多个科学包及其依赖包。Anaconda应用于多个系统，可以同时管理不同版本的Python环境，在本实验中使用的是Python3版本。

二、NLP概述

NLP既自然语言处理，指利用计算机对自然语言的形、音、义等信息进行处理，对字词句篇章的输入、识别、分析、理解、生成、输出等进行操作和加工的过程[3]。NLP的具体表现形式包括机器翻译、文本摘要、文本分类、文本校对、信息抽取、语音合成、语音识别等，在本文中，利用相关大数据技术在自然语言处理中的应用，研究NLP技术在文本中的情感分析。

（一）NLP基本流程

1.预料获取

（1）利用已经建好的数据集或第三方语料库。

（2）获取网络数据。

（3）制定数据搜集策略搜集数据。

（4）与第三方合作获取数据。

2.预料预处理

（1）去除数据中非文本内容。

通过语料获取文本内容后，文本数据中会存在很多无用的内容信息，如HTML、CSS、JSP以及标点符号等信息都需要去除，净化后的文本内容是相对纯文本信息。

（2）中文分词

去除数据中的非文本内容后，需要对长文本语句进行分词处理，常用的中文分词工具有jieba、FoolNLTK、HanLP、THULAC、NLPIR、LTP等，在本实验中使用jieba作为分词工具，jieba是使用Python语言编写的。

（3）词性标注

词性标注的任务是给词语打上标签，如名词、动词、副词、冠词、形容词等，常用的词性标注方法有基于规则的算法、基于统计的算法等。

（4）去停用词

中文文本中存在大量的虚词、代词或者没有特定含义的动词、名词，在文本分析的时候需要去掉。采用去停用词的技术，可以在减少句子中不必要词的情况下，保证原有句子的语义没有影响和意义变化。

3.文本向量化

文本数据经过预处理去除数据中非文本内容、中文分词、词性标注和去停用词后，基本上就是纯文本内容了。需要通过某些处理手段，预先将文本转化为特征向量才能参与任务计算。一般情况下可以调用模型对文本进行处理，如词袋模型、独热表示、TF-IDF表示、n元语法模型和Word2Vec模型等。

4.模型构建

NLP中使用的模型包括机器学习模型和深度学习模型两种。常用的机器学习模型有KNN、SVM、Naïve Bayes、决策树、K-means等。常用的深度学习模型有RNN、CNN、LSTM、Seq2Seq、FastText、TextCNN等。

5.模型训练

模型构建后，需要通过相应模型进行训练。训练模型时，为了避免训练时间过长，可以使用一部分数据进行实验。在模型训练中，会出现过拟合和欠拟合问题，解决好这两个问题，是模型稳定性的一个关键。同时也要避免梯度消失和梯度爆炸问题。

6.模型评价

模型训练工作完成后，需要对模型的效果进行评价。通过准确率、精确率、召回率、F1值、ROC曲线、AUC曲线等指标进行模型评价。

（二）语料库

NLTK即自然语言处理工具包，是用于处理自然语言数据的Python应用开源平台，NLTK提供了50多个素材库和词库资源的接口[5]，涵盖语料库获取、分词、词性标注、分类、聚类、分块、指标评测、概率和评估、命名实体识别、句法分析等多项NLP领域的功能，支持NLP和教学研究，收集的大量公开数据集和文本处理库可用于文本分类、符号化、贴标签、解析和语音推理等。

（三）关键词提取

文本是海量信息中最多且使用广泛的数据类型之一。在NLP领域中，从海量的文档中提取关键词汇，这些词汇能在一

定程度上体现文档的核心内容，进而实现查找内容的需求。

关键词能包括文本的主题，从而帮助阅读人员快速分析出所选的内容是不是感兴趣的内容。常见的关键词提取算法有TF-IDF算法、TextRank算法和LSA和LDA算法，其中LSA和LDA算法属于主题模型算法。

（四）文本向量化

随着NLP越来越多的应用机器学习和深度学习工具解决问题情况下，文本向量化成为NLP中非常重要的内容，因为文本向量化可将文本空间映射到一个向量空间，从而使文本可计算。一是计算机任何计算的前提都是向量化，而文本难以直接被向量化；二是文本的向量化应当尽可能地包含语言本身的信息；三是自然语言本身体现了人类社会的一种深层次的关系。

三、NLP技术在文本情感分析中的应用

情感分析技术的核心问题是情感分类，情感类别一般划分为正面、负面、中立三类或者正面、负面两类。另一种是包含悲伤、忧愁、快乐、兴奋四元分类，以及高兴、悲伤、喜欢、生气、厌恶、恐惧和惊讶七分类，同时也可以根据实际需要划分情感种类以及设置情感词。在本实验中，采用基于文本分类的方法进行文本情况分析。基于文本分类的方法采用标注了情感类别的文本进行训练，获取情感分类器，对情感分类器进行测试，输出含有多个概率值的结果，选择概率最高的情感倾向作为分类结果。

（一）特征选取

特征就是分类对象的所展现的部分特点，是实现分类的依据。需要根据实际情况选择有助于判断的特征，当文本量庞大时，特征量也会随之增大，信息量的增大无疑会影响运行速度，此时就需要对特征进行降维。可以采用统计词频、统计文档频率等方式进行特征降维。

第一步：获取关于流浪地球观后感的积极评论和消极评论，并作为两个语料库进行读取。

第二步：创建积极评论词典和消极评论词典。

第三步：对评论词典进行jieba分词。

第四步：对积极评论词赋予POSITIVE标签，对消极评论词赋予NEGATIVE标签。

（二）文本转换为特征变量

机器学习无法直接将中文文本作为输入数据直接处理，在本论文中使用分类算法将输入文本转换为特征向量的表现形式。True表示文本具有此特征，False表示不具有此特征。

如流浪地球观后感的积极评论词典中的一句“特效和美术满分”转换为特征，结果为[{“特效”：True，“和”：True，“美术”：True，“满分”：True}]。

（三）划分训练集与测试集

一个大的数据集处理时长较长，通常将积极和消极评论数据集分别划分训练集与测试集，训练集用于训练文本，测试集用于测试分类算法的效果。

划分数据集为训练集和测试集，一般按照二八比例，抽取80%数据作为训练集，抽取20%数据作为测试集。

（四）构建分类器

运用机器学习的算法训练数据集，得到分类器。选择机器学习算法时，可以根据实际情况调用合适的算法构建分类器，也可以同时采用多种算法，然后选用准确度最高的算法构建分类器。一般情况下，不同的文本所需要采用的分类器有所不同，所以需要采用多种算法进行训练，然后选用效果最佳的算法进行下一步测试。

采用nltk.classify的朴素贝叶斯模块的Train方法构建模型分类器，首先对训练集进行分类训练，然后用测试集对分类模型的效果进行准确率测验。

（五）验证分类器

分类器构建完成后，需要进行分类器的验证。使用测试集对分类器进行测试，比较测试结果，获得测试集的准确率，分析测试结果，给出改进建议。

根据上一步的训练出的模型分类器来对文本语句进行情感分析，首先读取要情感分析的文本句子并用True表示文本具有此特征，其次用训练出的模型的prob_classify函数对词语进行情感分类，分为消极和积极两类，最后，对每个分类结果进行正确率计算来验证是否合理。

本实验采用Python语言进行编写，主要对sentences = ['手感不错，推荐购买'，'破烂平板'，'外观不错，但是电池不耐用，差评'，'父母很喜欢，一次满意的购物'，'今天很开心'，'会回购']六句话进行了情感标签计算。

输出结果如下：

‘手感不错，推荐购买’的情绪面标签为 POSITIVE 概率为 97.11%;

‘破烂平板’的情绪面标签为 NEGATIVE 概率为 68.37%

‘外观不错，但是电池不耐用，差评’的情绪面标签为 NEGATIVE 概率为 83.38%

‘父母很喜欢，一次满意的购物’的情绪面标签为 POSITIVE 概率为 92.33%

‘今天很开心’的情绪面标签为 NEGATIVE 概率为 92.67%

‘会回购’的情绪面标签为 NEGATIVE 概率为 91.09%

准确率为：0.8432956381260097

四、总结

实验中主要对准确率、文本情感正负面情绪判断和概率值进行了计算。结果表示，测试数据的准确率约为84.33%，从信息量较大的前10个特征中发现大众对产品的评价不高，文本情感正负面情绪判断和概率值基本比较有效，分类器的结果较好。由于语料是来自只对某个产品的评价，因此适用的范围也只限制于相关评论内容的文本。如何增加全面语料库，在语料库全面的基础上进一步的提高测试数据的准确率，是下一步的研究方向。

参考文献

[1]段培吉，商思争，詹爱铃，易爱军，王亚光.人工智能时代我国会计人才需求分析——基于爬虫大数据分析技术[J].淮海工学院学报（人文社会科学版），2019，17（12）：78-81.

[2]陈斌.大数据支持下的网络日志分析技术研究[J].科技视界，2019（36）：19-20.

[3]王静宁，底慧萍.大数据背景下数据思维模式分析[J].卫星电视与宽带多媒体，2019（24）：58-59.

[4]李明，胡吉霞，侯琳娜，严峻.商品评论情感倾向性分析[J].计算机应用，2019，39（S2）：15-19.

[5]杨开漠，吴明芬，陈涛.广义文本情感分析综述[J].计算机应用，2019，39（S2）：6-14..

作者简介:

宋伟伟（1990-），女，汉族，系统分析师，硕士研究生，研究方向为大数据应用技术。