文章编号:1674-7046(2019)05-0074-06 DOI:10.14140/j.cnki.hncjxb.2019.05.013

基于文本情感分析的社交媒体数据挖掘

张 娜,柳运昌,王若男
(河南城建学院 计算机与数据科学学院,河南 平顶山 467036)

摘 要: 随着越来越多的社交媒体平台开始流行,针对社交媒体的信息挖掘与数据分析显得十分必要。对社交媒体进行评论数据爬取,对文本数据进行清洗预处理,去除停用词,然后采用贝叶斯定理对文本数据进行意见挖掘,最后根据获得的情感值生成饼图,对数据进行可视化呈现。

关键词: 文本挖掘; Python; 数据可视化; 数据分析
中图分类号: TP391.4 开放科学(资源服务)标识码(OSID):
文献标识码: A

Social media data mining based on text emotion analysis

ZHANG Na, LIU Yun-chang, WANG Ruo-nan
(*School of Computer & Data Science, Henan University of Urban Construction, Pingdingshan* 467036, *China*)

Abstract: As more and more social media platforms become popular, information mining and data analysis for social media are very necessary. Social media is crawled with comment data, these text data are cleaned and preprocessed to remove stop words, then Bayesian theorem is adopted to mine opinions on the text data, finally pie charts are generated according to the obtained emotion values, and the data are visually presented.

Key words: text mining; Python; data visualization; data analysis

文本情感分析又称为意见挖掘、倾向性分析等,是针对主观性文本进行挖掘、分析和推理的过程。随着互联网行业的飞速发展,各类信息数据爆炸式地增长,大数据得以出现。近年来,越来越多的社交媒体平台开始出现,如微信、微博、博客等,人们每天在这些社交软件上花费大量的时间,因此针对社交媒体的信息挖掘与数据分析十分必要。对微博这一社交平台进行文本数据挖掘,旨在了解人们对这些主观色彩的评论的看法以及人们的各种情感色彩和情感倾向性。

情感分析一词由 Hatzivassiloglou 等人首次提出,并采用情感词典使得文本情感分析的判断结果准

收稿日期:2019-05-28
基金项目:河南省科技攻关计划项目(172102210105);河南省重点研发与推广项目(182102210224,182402210025);河南省高等学校重点科研项目(18B520007,17A520024);平顶山市科技攻关项目(2017009(9.4))
作者简介:张 娜(1980-),女,河南商水人,博士,副教授,研究方向:人工智能、机器学习。

确率达到 82% [1]。Pang 等人采用支持向量机(SVM)、朴素贝叶斯(NB)在电影评论数据集上对评论文本进行情感判定[2]。在国内,张林等人以智能移动设备上发表的用户评论为研究对象,提出了一种基于短评论特征共现的特征筛选方法[3]。文献[4]提出一种基于性格的微博情感分析模型。唐晓波等人提出了一种基于旋进原则和 AdaBoost 集成技术的回归 SVM 情感分类模型,提高了主观文本标注的准备率,并实现了文本情感强度阈值的可视化[5]。黄磊等使用深度学习的神经网络模型对文本情感分析展开研究,以词向量作为基本输入单元,严格遵守单词之间的顺序,保留原文本中语义组合,克服了传统文本分类方法的不足[6]。

本文针对社交媒体进行评论数据爬取,介绍了网页数据的采集与挖掘、Python 的工作原理以及一些第三方库的使用,对爬取到的数据进行分词与关键词的提取、词云与词频统计以及文本情感分析。

1　文本情感分析

1.1　情感分类

情感分类是把评论归类为表达积极或消极情绪的任务。随着计算机计算能力的增强和数据量的爆炸式增长,传统机器学习已经满足不了人工智能快速发展的需要,终身机器学习(Lifelong Machine Learning)被学术界关注。终身机器学习具有快速学习、连续学习和迁移学习新任务的能力,在机器视觉、智能医疗诊断、搜索引擎、数据挖掘和机器人等领域具有应用价值。本文主要工作是采用终身机器学习的方法对情感分析进行分类。这类学习方法的思想是像人类一样学习,即根据过去学习的知识,形成数学模型,并通过知识推理帮助学习未知的知识,假定有 N 个学习任务 $T_1, T_2, \cdots, T_N$, $T_i = (f^{(i)}, X^{(i)}, y^{(i)})$,其中 $f^{(i)}$ 为真实隐含函数,它决定了每个任务的标签,$X^{(i)} \epsilon R^{d \times n_t}$, n_t 为给定的训练样例的个数,$y^{(i)}$ 为相应的标签。当学习器接收到学习任务时,该学习任务可能是一个已经学习过的任务也可能是一个新任务。接收到训练数据后,学习器会对已学任务进行预测,即在面临第 $N+1$ 个任务 T_{N+1} 时,学习器根据已知的训练样本可以学习每个任务的学习模型 $\tilde{f}^{(i)}$ 以接近真实的目标函数 $f^{(i)}$。通过所有测试样本上的误差决定学习性能。一个 *LL* 系统包含四个通用组件[7],如图 1 所示。

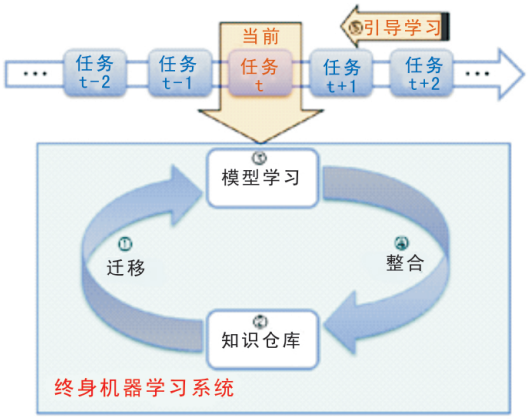

图 1　*LL* 系统框图

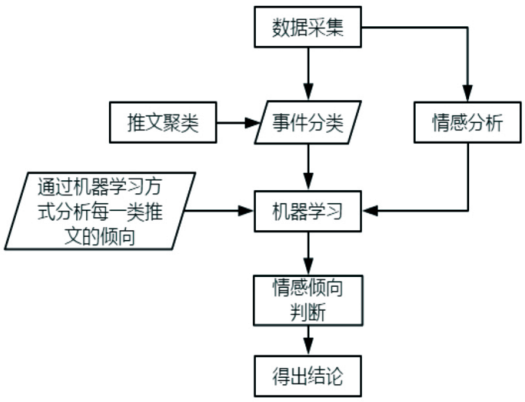

图 2　基于机器学习的文本情感分析流程

(1) 知识仓库: 主要用于存储过去经验中学习到的有必要存为后用的知识,机器学习系统是利用已经学习到的知识,因此知识仓库十分必要。目前知识仓库中的知识存储方式主要有两种: 一种是存储已学习的学习样本,缺点是造成空间的浪费; 另一种是保存知识的特征,也就是说要对原始信息特征进行提取和压缩。

(2) 过去的信息存储器: 该存储器存储过去经验中学习的知识。通常包括以下内容: 过去任务中的原始数据; 学习的中间结果; 基于过去的学习任务和学习数据得到的学习模型。

(3) 知识挖掘器: 根据过去的知识信息库学习获得信息并将知识存储在知识库中。

(4) 基于知识的学习过程: 包括迁移学习和整合知识,迁移学习是指选择知识仓库中对新模型有用的知识进行迁移,帮助新模型的学习,是终身学习系统的理论基础; 整合知识是在学习新任务的同时,对

知识进行筛选，去除无用的知识，同时兼顾整合后不会带来原有知识的损失。

本文将基于终身学习的方式用于文本情感分析，该终身学习的目标是执行一系列有监督的情感分类任务，每一个任务由训练文件组成，每个训练文件都带有正负标签，假定要学习第 N 个任务，需要从过去($N-1$) 个任务中获得的知识中学习更好的分类器。

1.2 文本情感分析基本流程

收集足够多的推文数据进行文本情感分析，采用推文聚类对所有的事件进行分类，完成分类后，进行基于知识的的学习过程，即对于分类结果的推文进行机器学习，模型学习在终身学习系统中负责新任务的快速学习，将学习得到的分类器运用于新的学习任务，即其他不可用情感分析得出分类结果的推文，同时将新的学习任务中学习到的新知识整合进原有的知识库中，得出最终结论。文本情感分析的基本流程如图 2 所示。

2 贝叶斯定理

朴素贝叶斯算法[4,7]：基于贝叶斯定理与特征条件独立假设的分类方法，首先介绍贝叶斯定理，然后介绍具体算法流程。贝叶斯定理可表示为:

$$p(c|x)=\frac{p(x,c)}{p(x)}=\frac{p(c)p(x|c)}{p(x)} \tag{1}$$

在文本情感分析中，x 为文本向量 $x=\{x_1,x_2,\cdots,x_d\}$，c 为文本对应的标记; $p(c)$ 是先验概率，表示某一类样本的存在概率; $p(c|x)$ 为后验概率，代表给定文本后，此文本属于哪一类的概率。

设文本向量 $x=(x_1,x_2,\cdots,x_d)$ 是 d 维向量，类标签 $y\in\{1,2,\cdots,n\}$; 给定文本向量 x，预测目标为寻找类别 k，使得公式(2) 最大。

$$p(y=k|x)=\frac{p(x|y=k)\cdot p(y=k)}{p(x)} \tag{2}$$

对于某一个样本 x，$p(x)$ 取值不变，所以预测 $\max\limits_k p(y=k|x)$，等价于预测 $\max\limits_k p(x|y=k)\cdot p(y=k)$ 的类别 k。

条件独立性假设: 对已知类别，所有属性相互独立。

条件独立性假设在于简化 $p(x|y=k)$ 的估计:

$$\begin{aligned}p(x|y=k)&=P(x_1,x_2,\cdots,x_d|y=k)\\&=p(x_1|y=k)\cdot p(x_2|y=k)\cdots p(x_d|y=k)\end{aligned} \tag{3}$$

基于条件独立性假设，贝叶斯算法等价于预测:

$$\arg\max_k p(x_1|y=k)\cdot p(x_2|y=k)\cdots p(x_d|y=k)\cdot p(y=k) \tag{4}$$

估计上述概率分布，采用极大似然估计，方法为:

$$p(y=k)=\frac{n_k}{n} \tag{5}$$

其中，n_k 表示标记为 k 的样本个数，n 为总的样本个数。

$$p(x_i=s|y=k)=\frac{n_{k,i_s}}{n_k} \tag{6}$$

其中 n_{k,i_s} 表示标记为 k 且文本向量的第 i 维的值为 s 的样本个数，n_k 表示标记为 k 的样本个数。

3 基于 NLP 的评论文本情感判断

NLP(Natural Language Processing) 即自然语言处理。微博、微信、博客等在线评论属于自然语言文本，人们往往用主观句来表达，一般都带有个人情感信息[5]。首先爬取训练数据，对数据进行预处理，采用贝叶斯算法实现文本情感判断。贝叶斯方法的文本情感分析操作流程如图 3 所示。

分类流程如下:

(1) 获取训练数据: 分别为积极评论和消极评论。

(2) 数据预处理: 对爬取到的文本数据进行文本分词,对停用词进行过滤,停用词(Stop Words)是指在处理自然语言数据时自动筛选出来的一些词语,这些词语往往不代表任何含义,甚至会妨碍数据挖掘的进行,首先对这些词进行清洗处理。使用向量来表示文本的内容,采用 TF - IDF 方法进行文本到向量的转化。TF - IDF(Term Frequency - Inverse Documents Frequency,即词频 - 逆文件频率)算法是一种统计方法,用于评估一个词语对于整个语料库的重要程度,即如果这个词语在整篇文档中出现的数量很高,而在所有的文档中出现的数量很低,就可以代表这篇文档。TF - IDF 算法如下:

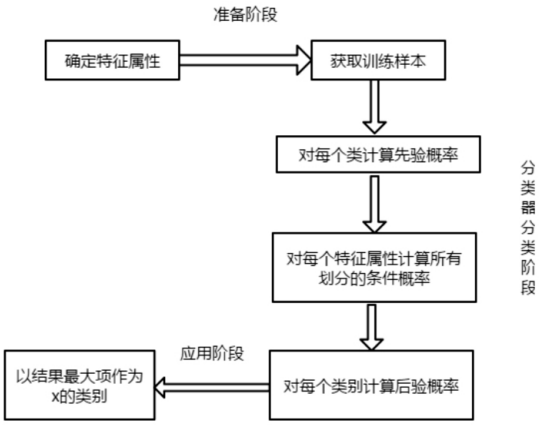

图 3　基于朴素贝叶斯情感分析流程

给定词典 $w=\{w_1,w_2,\cdots,w_v\}$,文档 d 为向量 $d=\{d_1,d_2,\cdots,d_v\}$,$tf(t,d)$ 表示词 t 在文档 d 中出现的频次,TF - IDF 方法在计算某个词的权重时,不仅考虑到词频,而且考虑到包含词的文档在这个文档集中的频次信息。计算公式为:

$$tf-idf(t)=tf(t,d)\cdot idf(t) \tag{7}$$

$$idf(t)=\frac{\log(n+1)}{df(t)+1}+1 \tag{8}$$

其中,$df(t)$ 代表文档集中词 t 出现的文档数量,n 表示所有的文档数目,$idf(t)$ 为此逆文档频率,它表征词在文档中的稀缺程度。

(3) 训练模型: 将文本向量和相应的文本标记输入到朴素贝叶斯算法中,估计概率 $p(y=k)=\frac{n_k}{n}$ 和 $p(x_i=s\mid y=k)=\frac{n_{k,i_s}}{n_k}$,其中 $i=1,2,\cdots,n$,n 表示字典的规模,$k=\{$积极,消极$\}$。

(4) 模型估计: 按照公式 $p(x\mid y=k_i)\cdot p(y=k_i)$ 第三步估计所得的概率,计算评论属于每一类的对应概率。

4　实验结果分析

4.1　文本数据预处理

文本数据预处理是在对所获取的数据进行处理之前,先对数据进行简单的调整,对这些数据选取分词处理,去除停用词以及关键词的提取工作。分词采用 Python 的分词工具——Jieba 分词,分词主要采用切割法: (1) jieba. cut: 输入两个参数,分别为需要进行分词的数据和使用 cut_all 参数控制是否使用全模式; (2) jieba. cut_for_research: 输入一个参数,即需要进行分词的数据,此方法适用于细粒度的搜索引擎的反向索引划分。

关键词的提取采用 TF - IDF 算法,基于该算法,通过 jieba. analyse. extract_tags 方法对于关键词的提取,该方法的参数见表 1。

表 1 TF – IDF 算法参数

参数	描述	备注
sentence	待提取文档	指定文档
topK	返回几个 TF – IDF 权重最大的关键词	默认 20
withWeight	仅包括指定词性的词	默认 False
allowPOS	仅包括指定词性的词	默认空

该程序获取的关键词如图 4 所示。

4.2 词云

词云，亦称文字云，是一种文本数据的呈现方式，是在文字数据里经常出现的“关键字”的视觉突出表现，形成一个关键字的渲染，输出一个云状的彩色图像，使文本数据的主要含义一目了然。Mask 翻译为蒙版，类似于 Photoshop 中的蒙版，蒙版即根据所提供的图片生成该形状的词云，对于图片的选择没有限制。在本设计中没有使用透明背景，选择使用蒙版。最终生成词云如图 5 所示。

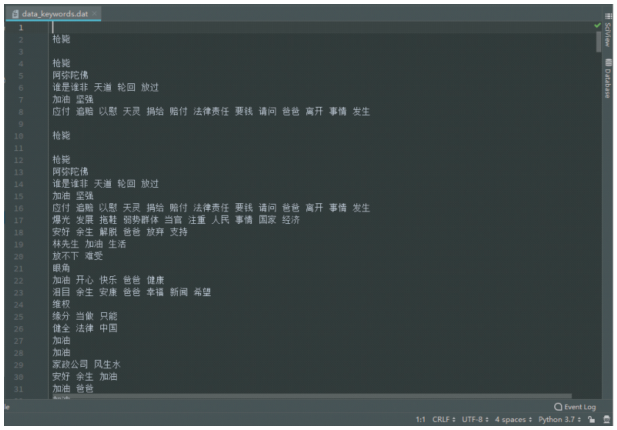

图 4 获取的关键词

图 5 生成词云

4.3 词频分析

词频分析(Word Frequency Analysis) ，顾名思义就是针对某些重要词语在整篇文档中的出现频率进行统计。本程序选用词频分析，针对新浪微博平台上众多网友的评论进行统计，生成的图片为条形统计图，如图 6 所示。其中，纵轴的刻度从 0 到 1，每一个刻度为 0.2，横轴的刻度是所提取的关键词频率，从左至右频率逐渐降低，图 6 选取了 20 个频率最高的关键词。

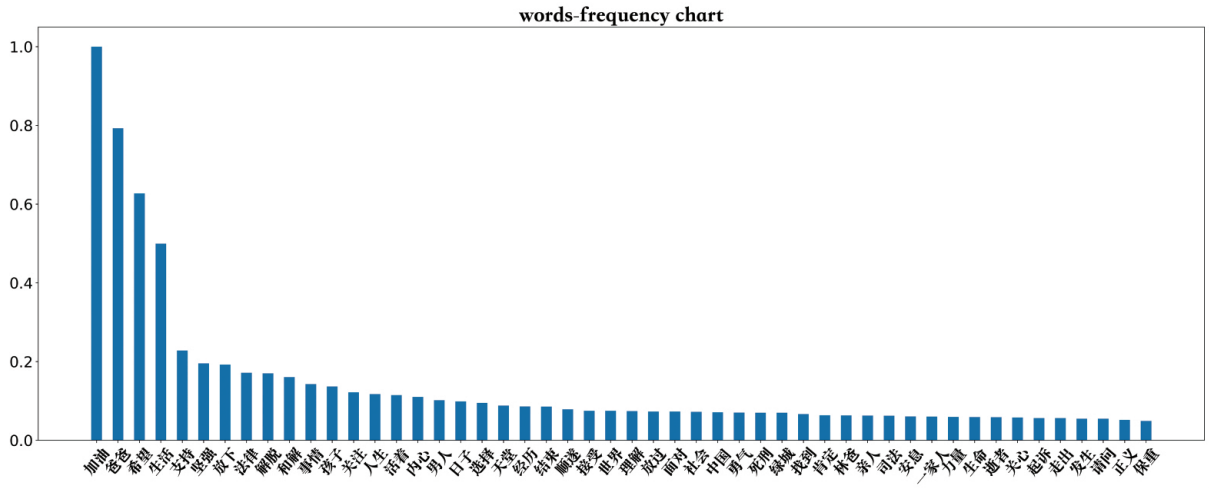

图 6 词频分析

4.4　文本情感分析

采用 Python 中的 snowNLP 类库对文本内容进行处理,该类库支持中文分词、词性标注、情感分析、文本分类、繁简转换、提取关键词、tokenization 和相似文本这些功能。在 snowNLP 库中的情绪判断使用 sentiments 命令,返回值若为负面情绪的概率,则更加接近 0; 返回值若为正面情绪的概率,则更加接近 1。

将爬取的数据信息分为两类: 积极和消极。其中,如果数据为 1,则判为积极,若为 -1,则判为消极。

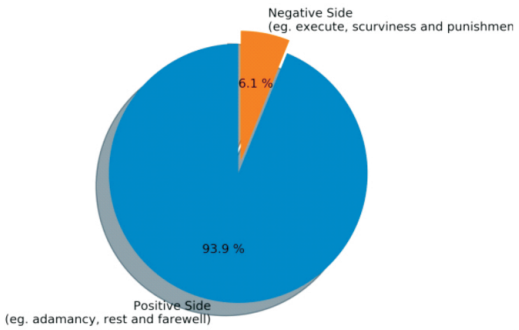

图 7　情感统计

朴素贝叶斯的程序代码逻辑[6]为: 首先对人工标注的类别进行计算,再经过文本的训练,将出现在语料库中的文本与所有文本统计,得出条件概率。使用测试文本乘词向量,最后得出该文本所属类别。

根据文字内容与对应数值,对经过处理所得到的情感值进行统计生成分布图,如图 7 所示。由图 7 可以看出 93.9% 的评论是相对积极的,只有 6.1% 的评论内容是负面的。

5　结束语

采用基于 Python 的网络爬虫技术对新浪微博评论进行爬取分析,得到了大量有效数据。经过清洗、词云分析、词频分析、情感分析,挖掘潜在信息。首先,对微博评论爬取后,将爬取到的数据及时写入数据库,再对这些文本数据进行清洗预处理,以去掉影响挖掘结果的词语,即停用词。情感分析又称为文本意见挖掘,是近来十分流行的一种对文本数据的主观情感色彩用以自然语言处理与计算机等方法进行分析处理的方法。本文采用 Python 中的 snowNLP 模块,基于概率贝叶斯定理,使用 snowNLP 库可以对文本数据进行意见挖掘,将文本数据归为积极和消极两类,最后根据所获得的情感值生成饼图,对数据进行可视化呈现。

参考文献

[1] Hatzivassiloglou V, Mckeown K R. Predicting the Semantic Orientation of Adjectives [C]. Proceeding of the Acl, 1997.
[2] 袁婷婷,杨文忠,仲丽君,等. 一种基于性格的微博情感分析模型 PLSTM [J]. 计算机应用研究,2018,37(2):1-7.
[3] Bo P, Lee L, Vaithyanathan S. Thumbs up: sentiment classification using machine learning techniques [C]. Proc of Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics. 2002: 79-86.
[4] 张林, 钱冠群, 樊卫国,等. 轻型评论的情感分析研究 [J]. 软件学报,2014, 25(12):2790-2807.
[5] 唐晓波,严承希. 基于旋进原则和支持向量机的文本情感分析研究 [J]. 情报理论与实践, 2013,36(1):98-103.
[6] 黄磊, 杜昌顺. 基于递归神经网络的文本分类研究 [J]. 北京化工大学学报: 自然科学版, 2017,44(1):100-106.
[7] 胡翔宇. 在线社交网络的用户倾向挖掘 [D]. 成都: 电子科技大学,2018.
[8] 孟天乐. 朴素贝叶斯在文本分类中的应用 [J]. 论述,2019,24(1):244-245.
[9] 刘玉林, 管利荣. 基于文本情感分析的电商在线评论数据挖掘 [J]. 统计与信息论坛,2018,33(12):119-124.
[10] 朱军,刘嘉勇,张腾飞,等. 基于情感词典和集成学习的情感极性分类方法 [J]. 计算机应用,2018,38(1):95-98.