文章编号：2095-2163(2021)03-0044-04 中图分类号：TP181 文献标志码：A

基于双层树状支持向量机的观点挖掘与倾向分析

孙 红[1,2]，黎铨祺[1]，赵 娜[1]

（1 上海理工大学 光电信息与计算机工程学院，上海 200093；2 上海现代光学系统重点实验室（上海理工大学），上海 200093）

摘 要：本文通过进行大量预处理工作，将经过词袋模型和 Word2Vec 两种不同向量化方法处理后的文本数据分别输入到 SVM 和 LSTM 模型中，训练出可以识别文本情感倾向的模型。进而对新产生的评论进行分类。根据实际数据量的倾斜状况，基于传统机器学习算法支持向量机（SVM），本文提出双层支持向量机，采用 2 种不同的方法分别训练模型并预测。最后再使用深度学习算法长短时记忆模型（LSTM）再次训练并预测，并对这 3 种方法做出比较和总结。结果显示，双层 SVM 比单层 SVM 的准确度提高了 8 个百分点；而 LSTM 比单层 SVM 低了 2 个百分点，比双层 SVM 低了接近 10 个百分点。

关键词：商品评论；网络爬虫；SVM；LSTM；情感分类；数据挖掘

View mining and trend analysis based on double-layer tree Support Vector Machine

SUN Hong[1,2], LI Quanqi[1], ZHAO Na[1]

(1 School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China; 2 Shanghai Key Lab of Modern Optical System(University of Shanghai for Science and Technology), Shanghai 200093, China)

【Abstract】 In this paper, a large amount of preprocessing work is carried out, and the text data processed by the following two different vectorization methods as the word bag model and Word2Vec are input into the SVM and LSTM models, respectively to train a model that can recognize the emotional tendency of the text. Further the newly generated comments are classified. According to the tilt of the actual data volume, based on support vector machine (SVM) that is the traditional machine learning algorithm, this paper proposes a two-layer support vector machine, using two different methods to train the model and predict. Thus, the deep learning algorithm long-term memory model (LSTM) is used to train and predict again, and the three methods are compared and summarized. The results show that the accuracy of the two-layer SVM is 8 percentage points higher than that of the single-layer SVM; while the LSTM is two percentage points lower than the single-layer SVM, which is nearly 10 percentage points lower than the double-layer SVM.

【Key words】 product reviews; Web crawler; SVM; LSTM; emotion classification; data mining

0 引 言

根据 2020 年 9 月第 47 次的《中国互联网络发展状况统计报告》[1]显示，截至 2020 年 6 月，国内网民规模达 9.40 亿，相较于上半年增长了 3 625 万，普及率达 67.0%，较 2020 年上半年提升 2.5 个百分点。互联网时代，人们普遍喜欢通过社交网络分享自己的生活和表达自己的观点，比如在朋友圈中表达日常生活中的快乐或者忧郁等情绪；在某个新闻 App 上发表自己对某件事情的看法；在购物网站上发表对某物品的使用感受。因此，在互联网中每天都会产生大量的用户评论，并且储存在互联网数据库中。如果能够充分地利用并挖掘这些信息，必然可以实现多种有效目的。但是，如果仅通过人工来对这些数据进行浏览和分析，则无疑会耗费大量人力资源，并且不能保证结果的准确性和可用性。这时就可以利用计算机强大的计算能力来帮助人们快速并准确地从这些海量主观性文本中分析出有用的信息，这就是文本的情感分析技术。

本文主要研究的是网购商品评论的情感分析技术，即从用户评论中通过文本挖掘技术提取信息。如果用户可以快速方便地从海量的主观文本中找寻到自己所需要的信息来指导自己的消费，那么对于用户的购物体验将会得到提升。

1 相关研究综述

1.1 国内外研究现状

情感分析最早由 Nasukawa 等人[2]提出。而文

作者简介：孙 红(1964-)，女，博士，副教授，硕士生导师，主要研究方向：大数据与云计算、控制科学与工程、模式识别与智能系统；黎铨祺(1994-)，男，硕士研究生，主要研究方向：机器学习、数据挖掘、深度学习；赵 娜(1992-)，女，硕士研究生，主要研究方向：机器学习、时序预测。

通讯作者：黎铨祺 Email：532166141@qq.com

收稿日期：2020-12-07

本的情感分析也叫文本意见挖掘或文本观点挖掘。更严格来说，两者的侧重点并不相同，文本意见挖掘根据给定的一段话中的文字或符号来判断这段话是趋向正面、还是负面。而文本观点挖掘更加偏重于理解这段文本真正的内在含义。

1.2　情感分析研究现状

本文最终定为文本意见挖掘，即判断目标文本表达了哪种情绪，分析后将情绪分为褒义、贬义两类；此外，一些比较复杂的分析则可以根据人的一般情绪来做区分，但从本质上来说都属于文本分类的任务。根据训练方式的不同，文本分类又可以分为有监督学习和无监督学习，对此拟做阐释分述如下。

（1）无监督学习。最大的特点在于不需要具有标签的数据集。所以，无监督学习可以减少大量繁琐的标注工作。Turney[3] 根据文本中的形容词或副词短语的平均语义倾向，对来自 4 个不同领域的文本进行聚类。陶娅芝[4] 使用基于 Word2Vec 的无监督方法对某个品牌手机的评论进行分类，避免大量的标注工作。

（2）有监督学习。需要大量已经标注好的数据，并且需要建立数学模型在这些标注好的数据中自动学习出数据的内在规律，从而根据这些内在规律完成情感分析任务。Pang 等人[5] 将朴素贝叶斯、最大熵分类和支持向量机用于电影评论的情感分类。

有监督学习往往需要用到已有标注好的语料进行训练，但是标注数据的获取却是一个较为繁琐的过程。而社交媒体网站就是一个天然的标注语料库，社交网络上的语料往往带有强烈的感情倾向，Bermingham 等人[6] 通过监测分析社交网络上公众对选举候选人的评论来预测政治选举的最终结果。韩萍等人[7] 使用一种基于自注意力机制的模型 E-DiSAN 来对社交网络评论文本的情感进行分类。但是，社交网站上通常没有用户的打分，只是一些带有感情色彩的主观性文本。而在这些文本中一般都夹杂着表达用户心情的特殊表情符号。崔安颀[8] 把特殊情感符号加入情感候选词库，作为其中一类情绪来进行情感分析。当然，如果采用这样的标注方法往往会伴随着许多噪声，Go 等人[9] 及 Pak 等人[10] 在远程监督的模型框架下，通过多重数据预处理，达到了去除噪声的效果。王义真等人[11] 利用 n-gram 的特性、词聚类的特征、词性标注的特征及否定的特征等构建出基于 SVM 的高维度混合特征算法模型，将其运用到短文本情感分类后，准确率得到了较大的提升。此外，还有许多应用于情感分析的方法，如 SVM[12]、依存句法[13]、卷积神经网络[14]、情感词典[15] 等。

2　数据预处理

从目标网站中爬取到的数据并不能直接放入模型中，需要对数据进行清洗与预处理。过程包括获取目标网站 URL、获取对应 Jason 页面、编写正则表达式、编写网络爬虫、循环爬取评论数据等。并将爬取得到的数据转化为可以输入模型的数据，具体步骤可分述如下。

步骤 1　替换和去除特殊符号。如果某个特殊符号与文本内容无关，则将其剔除；若其与文本内容有一定的关联，则选择一个通用词进行代替，比如遇到“666”、“6”、“耐斯”等词汇则使用“好”字将其代替。

步骤 2　繁转简。针对每个用户的输入法和地区的不同，某些评论可能会出现繁体字。

步骤 3　长句截断。由于传统支持向量机无法对超长句进行分析，这里将长句截断成短句。

步骤 4　中文分词。对上一个步骤截取的短句进行分词，并创建自定义词典。进行多次分词并筛选错误词汇加入自定义词表，最终得出一组比较完整的中文词。

步骤 5　将步骤 4 得到的词汇进行筛选，剔除出现次数不超过 5 次的词汇，保留剩余词汇作为词袋。词袋中根据每个词出现的次数将词按高到低进行，从 1 开始给每个词做上数字标记。

步骤 6　创建评论向量 ***numpy*** 矩阵，将步骤 4 得到的每条评论的词条与词袋中的词进行匹配，如果能匹配到，则用词袋词汇对应的数字编号来替代。最终得到一条条数字串评论向量，将所有的数字串评论向量进行拼接，限定长度，不足长度补 0，求得一个数字串评论向量组成的 ***numpy*** 矩阵。

3　建立分析模型与训练

3.1　支持向量机

支持向量机（Support Vector Machine，SVM）是 Cortes 等人[16] 在 20 世纪提出的用于解决分类问题的一种算法。SVM 的应用非常广泛，并已在多个领域取得研究成果。石强强等人[17] 通过增加情感词典的种类、提高系统对网络新兴词汇和特殊表情符号的识别，使用支持向量机模型对某些酒店的网站评论进行情感分类。郝晓燕等人[18] 分别使用支持

向量机算法、KNN 算法和最大熵模型进行了基于特征词布尔值的中文文本分类实验。

一个普通的 SVM 就是一条普通直线,这条直线用来完美划分线性可分问题的 2 个类别,如图 1 所示。

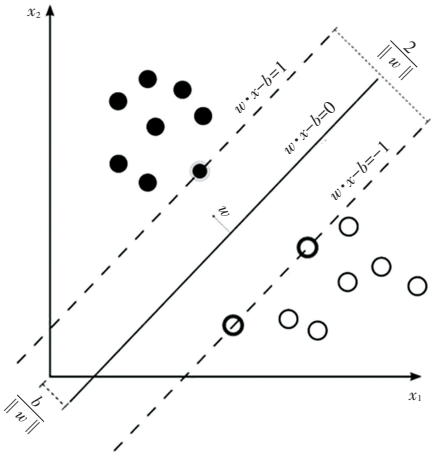

图 1 支持向量机

Fig. 1 Support Vector Machine

这里,首先定义直线 $y(x) = w^{\mathrm{T}}x + b$,则任意点 x_0 到该直线的距离如式(1)所示:

$$\frac{1}{\|w\|}(w^{\mathrm{T}}x_0 + b), \tag{1}$$

然后,对式(1)进行归一化运算,使得训练集 (x_i, y_i), $i = 1, \cdots, n, x \in R^m$ 满足式(2):

$$y_i(w^{\mathrm{T}}x_i + b) - 1 \geqslant 0, \quad i = 1, \cdots, n. \tag{2}$$

这时候,分界线到两边两类的距离等于 $2/\|w^2\|$,而最终的目标是,让分界线到两边的距离最大化,这就相当于最小化 $\|w^2\|$,如此就得到最优分类面。对于 N 个训练点的信息 (x_i, y_i),也可以写成如下数学形式:

$$\arg\max\left\{\frac{1}{\|w\|}\min_n[y_i(w^{\mathrm{T}}x_i + b)]\right\}, \tag{3}$$

虽然目标函数可以表达得很清楚,但实际上很难计算。通过引入利用拉格朗日乘子法,便可以通过一系列数学运算得到最终的目标函数,其计算公式为:

$$\max_{a//a_n} L(a) = \sum_{n=1}^{N} a_n - \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} a_n a_m y_m y_n x_n^{\mathrm{T}} x_m, \tag{4}$$

约束条件为式(5),即:

$$a_n \geqslant 0, \forall n, \sum_{n=1}^{N} a_n y_n. \tag{5}$$

通过引入核技巧将低维数据映射到高维空间可以提升模型的效果。类似于这种将某个特征空间的向量映射到另一个特征空间的函数就称为核函数[16],由于在 SVM 优化中,所有的运算表达都是内积,所以,这里可以把内积运算过程替换成核函数,从而不必做优化运算。

3.2 双层树状 SVM

对单层普通的支持向量机,结果显示分类效果并不明显。对数据进行分析得出,原因是数据倾斜非常严重,爬取的数据包含的正、负、中性评论分布严重不均匀。正向评论数量为 12 000 条,中性评论数量为 2 000 条,负向评论数量为 6 000 条。

为了能够有效缓解数据倾斜所带来的问题,本文提出双层支持向量机的方法,原理如图 2 所示。

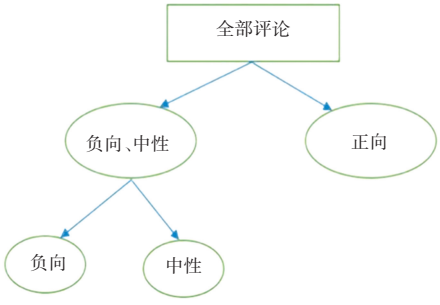

图 2 双层树状 SVM 原理图

Fig. 2 Schematic diagram of double-layer tree SVM

图 2 中,首先将中性和负向评论作为一类,与正向评论进行划分。再对中性和负向评论进行划分。这样在理论上就将数据倾斜带来的误差降低到最小。

先将中性和负向评论的标签置为 0,与正向评论的标签 1 相区分。处理好的数据作为总的数据输入,步骤同单层支持向量机,引入 KFold 划分数据,训练模型,验证模型。

4 结果对比与分析

设置好超参数后,使用之前分批处理过的京东商城和淘宝网的评论语料文本分别进行训练和测试,得到数据见表 1。

表 1 实验结果对比

Tab. 1 Comparison of experimental results %

方法	准确率	召回率	F_1 值
SVM	81.30	80.53	81.98
Tree-SVM	89.78	90.20	89.90
LSTM	79.46	80.20	78.63

由表 1 的结果可以看出:双层 Tree-SVM 表现效果最好,目前热门的循环神经网络的表现要逊色

于普通 SVM。究其原因，分析后可知：

首先，普通 SVM 在分类性能上已经相对比较成熟，对于这些特征明显，特征数量众多的文本，则能做出很好的区分。

其次，双层 Tree-SVM 是专门针对这个实验数据集的特征（三分类数据分布不均，正向评论数量远远大于负向和中性评论的数量）而产生的。所以，能在普通 SVM 的基础上，更好地切合这个数据集，从而表现出更佳的性能。

5　结束语

本文首先分析了 Web 2.0 时代的到来对当今社会产生的冲击，以及网络数据的发展态势。然后，提出核心技术：情感分析技术。简单介绍了部分经典以及当下流行的几种情感分析的算法模型。进而，分析数据获取的方式，提出网络爬虫的概念，介绍几种不同的网络爬虫框架，并分析爬取过程中可能出现的问题以及解决方法；根据实际情况编写 2 套分别适用京东和天猫的网络爬虫，循环爬取网站评论数据，进行分批式存储。在此基础上，分析爬取的数据，总结规律，根据实际数据情况，提出方法：普通支持向量机、双层树状支持向量机（Tree-SVM）和长短时记忆模型（LSTM）。最后清洗数据，主要包括中文分词、去停用词、文本向量化等，将数据输入进算法模型进行训练并验证。通过多次训练和验证，双层树状 SVM 在准确率上表现为 89.78%，与普通 SVM 相比高出 8 个百分点；而 LSTM 的准确率仅为 79.46%，但这并不能表示 LSTM 在性能上就不如传统机器学习方法，分析原因可能是数据量的不足，造成神经网络未能有效训练。

关于分词方面，本文使用结巴分词默认的通用词典，而对于一些手机评论中特有的词语，比如“吃鸡”、“打王者”、“王者荣耀”等则需要自行手动添加进去，由于研究时间有限，难免会有遗漏，而结巴分词的新词识别功能也只对 2 个字的词语有效果。需要构建出一个针对电子产品的用户字典，更加准确地分词。再比如一些网络上最近才出现的新兴词汇：“马甲”、“水友”、“水军”、“带躺”、“躺赢”等等，这些词往往具有很强的情感倾向，在今后的分析中可以做更进一步改进。

参考文献

[1] 中国互联网络信息中心. 第 46 次中国互联网络发展状况统计报告[R]. 北京：中共中央网络安全和信息化委员会办公室，2020.

[2] YI J, NASUKAWA T, BUNESCU R, et al. Sentiment analyzer: extracting sentiments about a given topic using natural language processing techniques [C]//Third IEEE International Conference on Data Mining. Melbourne, FL, USA: IEEE, 2003: 427-434.

[3] TURNEY P D. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews [C]// Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002: 417-424 .

[4] 陶娅芝. 基于 word2vec 和自训练的无监督情感分类方法[J]. 科技风，2019(12)：92-93.

[5] PANG B, LEE L, VAITHYANATHAN S. Thumbs up? Sentiment classification using machine learning techniques [C]// Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10. Association for Computational Linguistics. New York: Association for Computational Linguistics, 2002: 79-86.

[6] BERMINGHAM A, SMEATON A. On using Twitter to monitor political sentiment and predict election results[C]// Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2011). Chiang Mai, Thailand: Asian Federation of Natural Language Processing, 2011: 2-10.

[7] 韩萍，孙佳慧，方澄，等. 基于情感融合和多维自注意力机制的微博文本情感分析[J]. 计算机应用，2019，39(S1)：75-78.

[8] 崔安颀. 微博热点事件的公众情感分析研究[D]. 北京：清华大学，2013.

[9] GO A, BHAYANI R, HUANG L. Twitter sentiment classification using distant supervision[R]. CS224n Project Report, Stanford: Digital Library Technologies Project, 2009.

[10] PAK A, PAROUBEK P. Twitter as a corpus for sentiment analysis and opinion mining[C]// International Conference on Language Resources and Evaluation (Lrec 2010). Valletta, Malta: dblp, 2010: 1320-1326.

[11] 王义真，郑啸，后盾，等. 基于 SVM 的高维混合特征短文本情感分类[J]. 计算机技术与发展，2018，28(2)：88-93.

[12] 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，2020，43(8)：112-119.

[13] 梁晓敏，徐健. 舆情事件中评论对象的情感分析及其关系网络研究[J]. 情报科学，2018，36(2)：37-42.

[14] 陆敬筠，龚玉. 基于自注意力的扩展卷积神经网络情感分类[J]. 计算机工程与设计，2020，41(6)：1645-1651.

[15] 安璐，吴林. 融合主题与情感特征的突发事件微博舆情演化分析[J]. 图书情报工作，2017(15)：120-129.

[16] BENNETTK, DENIRIZ A. semi-supervised support vector machines [C]//Advances in Neural Information processing systems. Denver, Colo, USA: The MIT Press, 1999, 2: 368-374.

[17] 石强强，赵应丁，杨红云. 基于 SVM 的酒店客户评论情感分析[J]. 计算机与现代化，2017，17(3)：117-121.

[18] 郝晓燕，常晓明. 中文文本分类研究[J]. 太原理工大学学报，2006，37(6)：710-713.

[19] HUANG Chenghui, YIN Jian, HOU Fang. A text similarity measurement combining word semantic information with TF-IDF method[J]. Chinese Journal of Computers, 2011, 34(5): 856-864.