

结合扩充词典与自监督学习的网络评论情感分类

景 丽 李曼曼 何婷婷

河南财经政法大学计算机与信息工程学院 郑州 450046

(1286041400@qq.com)

摘 要 在高速发展的互联网时代,网络评论情感分析对分析舆情、监控电商有着重要作用。现有分类方法主要有情感词典方法和机器学习方法。情感词典方法过于依赖词典中的情感词,情感词典越完备,网络评论情感倾向越显著,分类效果越好,但对于情感倾向不易区分的评论,其分类效果欠佳。机器学习方法是一种有监督的方法,其分类效果依赖于大量事先标注的语料,目前语料标注是通过人工完成,工作量极大。文中综合了情感词典和机器学习两种方法的特点,构建了一个网络评论情感分类模型,利用相关领域网络评论对情感词典进行扩充,基于情感词典方法的分类结果,通过自监督学习训练一个分类器,进而提高情感倾向模糊文本的分类正确率。实验表明,与情感词典方法和机器学习方法相比,所提模型在酒店评论、京东评论两个数据集上都获得了更好的情感分类效果。

关键词: 网络评论;情感分类;词向量;情感词典;机器学习

中图法分类号 TP391.1

Sentiment Classification of Network Reviews Combining Extended Dictionary and Self-supervised Learning

JING Li, LI Man-man and HE Ting-ting

School of Computer and Information Engineering, Henan University of Economics and Law, Zhengzhou 450046, China

Abstract In the rapidly developing Internet era, sentiment analysis of online reviews plays an important role in analyzing public opinion and monitoring e-commerce. Existing classification methods mainly include sentiment dictionary methods and machine learning methods. The sentiment dictionary method relies too much on the sentiment words in the dictionary. The more complete the sentiment dictionary, the more pronounced the sentiment tendency of online comments and the better classification effect. The classification effect of comments is not good when the sentiment tendencies are not easy to distinguish. The machine learning method is a supervised method, and its classification effect relies on a large number of pre-annotated corpora. Currently, the corpus annotation is done manually, and the workload is extremely large. This paper combines characteristics of the two methods to build a new sentiment classification model of network reviews. First, the sentiment dictionary is expanded based on the domain of online reviews, and the sentiment value of each online comment is calculated according to the extended sentiment dictionary. According to the preset sentiment threshold, the comments with significant sentiment tendencies and higher accuracy are selected as the definite set, and the rest that are not easily distinguished are used as uncertain sets. The classification result of the definite set is directly determined by the sentiment value. Second, according to the definite set from the sentiment dictionary method, a classifier is trained through self-supervised learning, and the training data do not require manual annotation. Finally, the trained classifier is used to classify the uncertain set again, and an improved algorithm is used to improve the classification result of the uncertain set. Experiments show that, compared with the sentiment dictionary method and the machine learning method, the proposed model achieves a better sentiment classification effect for the sentiment classification of hotel reviews and Jingdong reviews.

Keywords Internet reviews, Sentiment classification, Word vectors, Sentiment dictionary, Machine learning

1 引言

互联网的快速普及改变了人们的的生活方式,人们不仅是信息的被动接受者,也是社会信息的主动生产者。越来越多的用户在网络上表达自己的看法,比如在微博平台上发表对某个事件的看法,在电商网站上发表对某个产品的评价。这

些网络评论蕴含着丰富的感情色彩(如积极、消极、赞同、反对等),也蕴含着巨大的商业价值和社会价值。因此通过相应技术进行情感分析,挖掘网络评论中的有价值信息,是时下的研究热点。

文本情感分析是指对蕴含用户情感的主观性文本进行分析、处理和总结,涉及自然语言处理、数据挖掘等多个研究领

基金项目:国家自然科学基金(61806073,31700858,61802110)

This work was supported by the National Natural Science Foundation of China(61806073,31700858,61802110).

通信作者:李曼曼(lmm18037193250@163.com)

域^[1]。目前的研究方法主要有两种:基于情感词典的分类方法和基于机器学习的分类方法。基于情感词典的分类方法中,情感词典越完备,网络评论的情感倾向越显著,分类的准确率越高;机器学习情感分类方法的准确率较高,但是需要充足且正确标注的语料来训练分类器。

相关研究表明,情感词典方法和机器学习方法是可以互补的^[2],情感词典方法对于消极文本分类正确率较高,机器学习方法对于积极文本分类正确率较高。经实验发现,该结论对于本文数据集也是成立的。基于以上研究发现,本文融合情感词典和机器学习两种方法的优缺点构建分类模型,通过对情感词典进行领域知识词汇扩充,用情感词典方法对网络评论进行分类,通过情感词典分类结果获取训练集并训练机器学习分类器,再利用机器学习分类器改进情感词典的分类结果。该模型有效解决了机器学习方法的标注语料问题和情感词典方法对情感倾向模糊评论不易分类的问题。实验表明,与情感词典方法和机器学习方法相比,该模型在酒店评论、京东评论的情感分类上均取得了更好的情感分类效果。本文的贡献主要有以下3个方面:

(1)领域情感词典的扩展。情感词典的覆盖程度对情感分类的正确率起到决定性作用,本文将 SO-PMI 与 word2vec 结合进行领域情感词典的扩充,这样既考虑了词汇的语义信息,又有助于提升情感分类效果。

(2)通过自动标注语料实现自监督学习。本文用词典和规则计算每个句子的情感值,通过设定情感阈值确定机器学习分类器的训练集,根据情感值自动标注语料,不需要人工标注。

(3)结合情感词典和机器学习方法修正不易区分情感倾向的网络评论。研究发现,机器学习方法对积极文本的分类效果较好,情感词典方法对消极文本的分类效果较好,因此本文结合两种方法对不易区分情感倾向的网络评论分类结果进行修正。

2 相关工作

基于情感词典方法的研究方向主要是情感词典的构建,通过提升情感词典的质量来提升情感判别效果。情感词典最早可追溯到 1997 年,Hatzivassiloglou 等^[3]从华尔街日报数据中挖掘出许多形容词词性的情感词语。之后,国内外学者不断对词典进行完善。Wiebe^[4]承继相似的思想,结合语料库,基于相似性聚类获得了形容词词性的情感词语。为了避免词性限制,Turney 等^[5]提出了点互信息算法,通过计算词语与种子词之间的互信息来识别某个词语是否为情感词,这种方法突破了词性限制,能够识别各种词性的情感词语。Li 等^[6]在英文词典的基础上,借助机器翻译,利用标签传播算法衡量词汇的情感信息,构建了中文情感词典。上述情感词典是最基础的情感词典,覆盖程度比较低,后来研究者对情感词典做了进一步的完善。Wang 等^[7]在 40 万条微博评论的基础上,利用统计信息和点互信息算法扩充情感词典,并根据微博特性,以表情符号作为附加信息提供辅助作用。Fan 等^[8]通过互信息来构建电影评论的领域情感词典,统计出评论中的高频、正负核心词共 30 个,用互信息计算核心词和新词之间的相似度,进而挖掘出电影领域的情感词。Radovano 等^[9]设置了两个基础类别词:“excellent”和“poor”,“excellent”代表积极,“poor”代表消极,并且假定与这两个词语一同出现的

词汇具有与它们同样的情感倾向,此种方法通过联系相关词汇的上下文信息来预测词汇的情感倾向。Jha 等^[10]通过计算不同词汇之间的共现信息来衡量词语之间的关联,并且结合基准词的极性估计未知词语的情感极性,进而对评论的情感倾向性进行分析。

机器学习分类方法的分类性能与提取的数据特征有关。Pang 等^[11]最早利用电影评论为数据,使用 N 元语法、词的位置以及词性作为文本分类的特征,选择使用朴素贝叶斯、最大熵、支持向量机 3 种不同的算法训练分类器来判断文本情感倾向,实验表明支持向量机分类器的效果最好。Paltoglou 等^[12]考虑到不同特征对类别判定的作用有差别,使用词频-逆文档频率方法计算文本中词语的特征权值,然后将词语的权重信息输入机器模型,实现对文本的情感分类。Tripathy 等^[13]将文本分词之后,考虑了上下文词语的联系,以 1 个词语、2 个词语、3 个词语及其组合的方式划分,然后分别采用朴素贝叶斯、最大熵、随机梯度下降、支持向量机方法进行评论情感倾向性分析。Mikolov 等^[14]于 2013 年推出 word2vec 模型,该模型能够将词语转化为空间向量,为特征提取开辟了新的思路。Wei 等^[15]利用 word2vec 模型提取电商产品评论文本特征,并通过词频-逆文档频率进行加权,最后采用支持向量机分类器进行分类,结果表明在情感极性分类问题中,加权后的词向量相较于传统特征提取方法对模型的提升有着明显的正向效果。Wang 等^[16]针对现有的情感词典不能很好地覆盖微博文本中情感词的不足,结合特定的情感符号和基于卡方统计量的度量方法,实现对情感词典的扩充,引入 word2vec 模型实现情感词和其所在微博语句的向量化表达,并利用 KNN 分类器实现微博句子级的情感分类。

3 本文模型

本文提出了一种融合情感词典和机器学习的情感分类模型,其流程如图 1 所示。模型分为两个步骤:1)基于扩充的情感词典计算网络评论情感值,根据预设情感阈值选取具有显著情感倾向和较高准确率的评论为确定集,剩下的情感倾向不易区分的评论为不确定集;2)以确定集为训练集,利用机器学习算法训练分类器,再利用训练过的分类器对不确定集进行分类,并用改进算法改进不确定集的分类结果。

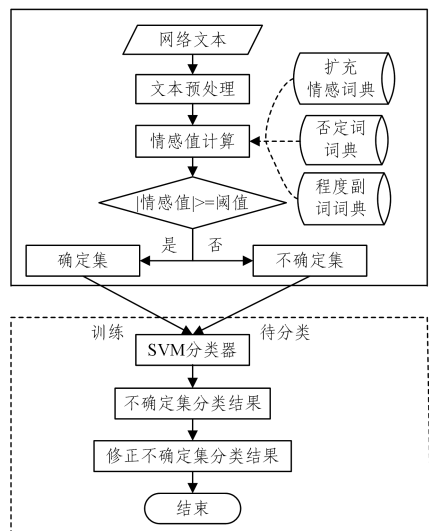


图 1 融合情感词典和机器学习的分类模型流程图

Fig. 1 Flow chart of classification model combining emotion dictionary and machine learning

3.1 扩充情感词典

文本的情感倾向由包含的情感词所体现,国内的学者已经整理了一些质量很高的情感词典。本文在其他研究的基础上,整合了知网的 HowNet 情感词典与台湾大学的 NTUSD 情感词典,最终得到积极词汇 6060 个,消极词汇 10270 个,程度副词 215 个,否定词 42 个,作为本文的基础情感词典。

网络评论涉及的领域较多,而上述情感词典含有的情感词语有限,如果只通过该词典对网络评论文本进行分类效果欠佳。为了识别网络评论领域的专有情感词,本文采用 word2vec 和 SO-PMI 相结合的方式扩充情感词典,该方法与传统的 SO-PMI 相比,能够考虑情感词的语义信息。

结合 word2vec 和 SO-PMI 扩充领域情感词典,需要人为选出高频且情感倾向明显的积极基准词和消极基准词,再将未知情感倾向的领域候选词与基准词中的情感词作相应运算,衡量领域候选词与基准词的相似度。如果相似度较高,依据基准词的情感倾向判定领域候选词的情感倾向,进而扩充网络评论涉及的领域情感词。

利用 word2vec 模型使情感词向量化,不同词语之间的语义相似度利用向量的余弦值来计算,具体计算公式如下:

$$SD(word_1, word_2) = \frac{vector(word_1) \times vector(word_2)}{\|vector(word_1)\| \times \|vector(word_2)\|} \tag{1}$$

其中,word₁ 和 word₂ 是词向量,|| word₁ || 与 || word₂ || 是两个词向量的模。SD(word₁, word₂) 是 word₁ 和 word₂ 之间的相似距离,其值越大则两个词汇在语义上越相似,其值越小则两个词汇越不相关。

基于式(1),word 的 SO-PMI 值的具体计算公式如下:

$$SO-PMI(word) = \frac{\sum_{pword \in Pwords} SD(word, pword) - \sum_{nword \in Nwords} SD(word, nword)}{2} \tag{2}$$

其中,Pwords 是积极基准词词典,pword 是其中的积极基准词;Nwords 是消极基准词词典,nword 是其中的消极基准词。

如果 SO-PMI 值大于 0 则 word 为积极情感词,如果 SO-PMI 值等于 0 则 word 不作为领域情感词,如果 SO-PMI 值小于 0 则 word 为消极情感词。经过该方法,在酒店评论中扩充了 532 个消极词汇,261 个积极词汇;在京东评论中扩充了 516 个消极词汇,115 个积极词汇。

3.2 情感值计算

在计算网络评论情感值时,需要先依据网络评论中的标点符号把评论文本切分成若干个分句,网络评论就由这些分句构成。先计算分句的情感值,将分句的情感值相加得到整条网络评论的情感值。

根据标点符号把每一条网络评论分成 m 个分句,根据分句中出现的情感词、程度副词、否定词计算情感值,计算公式如式(3)所示:

$$S = \sum_{i=1}^m \sum_{j=1}^n a_i \times w_{ij} \tag{3}$$

其中,S 表示网络评论的情感值,a_i 代表第 i 个短句中情感词前程度副词或者否定词的权重,w_{ij} 表示第 i 个短句中的第 j 个情感词的情感值。

使用情感词典、程度副词词典、否定词词典计算情感值时需要注意以下几点:

(1)情感词典。计算网络评论情感值时,使用情感词典与评论中的情感词进行匹配,设积极情感词的情感值为 1,消极

情感词的情感值为-1,然后求和并计算出评论的情感值。

(2)程度副词词典。副词本身没有感情色彩,但如果用于修饰情感词语,能够增强网络评论的情感程度。在网络评论中匹配出情感词之后,依据程度副词词典查找该情感词与前一个情感词之间是否存在程度副词,若存在程度副词,则该情感词的情感值乘以副词对应的权重值。

(3)否定词词典。判断是否存在程度副词之后,再判断情感词前是否出现否定词,如果情感词前有否定词,统计否定词的数量,若有偶数个否定词,则情感词的情感倾向没有反转,若有奇数个否定词,则进行反转取值操作^[17]。

3.3 确定集和不确定集的筛选

根据计算出的评论情感值,对评论进行积极和消极二元分类。研究表明,一条网络评论的情感值的绝对值越大,该条评论的情感倾向越显著,分类越准确,情感值的绝对值越小,情感倾向越不易区分。本模型预设一个情感阈值 D,依据情感值的绝对值大小把网络评论分为确定集和不确定集,具体划分方法如算法 1 所示。

算法 1 确定集和不确定集的划分

```
if |Si| ≥ D
    Ci ∈ {确定集}
else
    Ci ∈ {不确定集}
```

算法 1 中,S_i 是网络评论 C_i 的情感值,D 是情感阈值。

情感阈值 D 把评论分为两个集合,情感值的绝对值大于或等于 D 的评论属于确定集,小于 D 的评论属于不确定集。其中,在确定集中情感值大于或等于 D 的评论是积极类,情感值取反大于或等于 D 的评论是消极类。确定集用来训练机器学习分类器,然后利用经过训练的分类器对不确定集的分类结果进行修正。

情感阈值 D 的选取依据确定集中的评论数量和正确率,但是二者相互矛盾,D 值较大时,确定集中的评论数量少但正确率高;相反,D 值较小时,确定集中的评论数量多但正确率低。为了更好地训练机器学习分类模型,既要保证数据量充分又要保证正确率高,因此在选取情感阈值 D 时,针对确定集中的评论数量和正确率进行折中考虑。

3.4 修正不确定集的分类结果

本文将情感词典分类结果中的确定集作为训练集训练机器学习分类器,然后利用该分类器对不确定集进行分类,最后结合情感值和分类器的分类结果对不确定集的判定结果进行修正。具体修正方法如算法 2 所示。

算法 2 不确定集的判定结果修正

```
if Sj < 0
    Cj ∈ {消极评论类}
else
    if Dj = 'neg'
        Cj ∈ {消极评论类}
    else
        Cj ∈ {积极评论类}
```

算法 2 中,D_j 为不确定集中第 j 条评论的机器学习分类器的分类结果,D_j = 'neg' 表示分类结果为消极类,D_j = 'pos' 表示分类结果为积极类。对于不确定集中的第 j 条评论 C_j,如果其情感值 S_j < 0,不管机器学习分类器的分类结果如何,把 C_j 划分到消极倾向类中;如果其情感值 S_j > 0, C_j 的归属取决于机器学习分类器的结果。这是因为基于情感词

典的方法对消极评论的分类效果较好,而机器学习方法对积极评论的分类效果较好。

4 实验

4.1 实验数据集

本文实验使用的数据集共有 4 个:1)wiki 中文,wiki 官网下载,质量比较高并且涉及的领域十分广泛;2)微博数据,来自于 aitechclub 网站公开的资源,共 10 万条微博评论;3)酒店评论,该评论由中科院谭松波教授整理^[18],是一个二分类公开中文数据集,语料规模为 1 万条,本文实验选用其中正负类语料各 3 000 条;4)京东评论,在京东商城上爬取掌上电脑评论,将四星和五星标注为积极评论,将一星和二星标注为消极评论,积极和消极评论各标注 1 500 条。

wiki 中文和微博数据是训练词向量模型所使用的语料库,酒店评论和京东评论是情感分析实验所使用的语料库。wiki 中文涉及的领域广泛,微博数据适应了口语化比较严重的网络数据,这两个数据集训练出来的词向量模型能够获取不同领域的网络评论向量化表达。

4.2 数据预处理

实验使用结巴分词、中科院计算所发布的停用词表对数据预处理。用 10 万条微博和 wiki 中文数据集预训练 word2vec 模型,训练参数设置如表 1 所列。

表 1 词向量模型训练参数表

Table 1 Word vector model training parameter		
参数	参数值	意义
-sg	1	Skip-Gram 语言模型
-size	300	词向量维度
-window	5	词向量上下文最大距离
-alpha	0.000 1	学习速率
-min_cout	5	需要计算词向量的最小词频
iter	5	迭代次数

机器学习方法中,网络评论的向量化表达由 word2vec 模型实现,维度为 300,采用 PCA 算法降维到 100,实验选用 100 维作为下文机器学习分类算法的输入。

4.3 评价指标与情感阈值

本文使用的评价指标有精确率(Precision)、召回率(Recall)、F1、正确率(Accuracy)。

本文模型中的情感阈值 D 依据实验设定。对酒店评论和京东评论两个数据集 9 000 条评论计算情感值,表 2 列出了选取不同阈值时确定集和不确定集的评论数量和正确率。情感阈值越大,确定集的正确率越高,但评论数量越少。为了使本文模型获得较好的分类效果,需要折中考虑确定集评论数量和正确率。当情感阈值为 4 时,确定集正确率虽然比阈值为 5 时低 1.16%,但是确定集的评论数量多出 431 条,占总评论数据集的 4.78%。综合考虑,本文选取情感阈值为 4。

表 2 不同阈值对分组结果的影响

Table 2 Effect of different thresholds on grouping results				
D	确定集		不确定集	
	评论数	正确率/%	评论数	正确率/%
1	7 248	85.78	1 752	63.98
2	7 077	87.07	1 923	61.15
3	6 885	88.82	2 115	57.79
4	6 636	90.40	2 364	56.64
5	6 205	91.56	2 795	59.28

4.4 实验结果与分析

将本文模型与传统情感词典方法和机器学习方法的情感分类正确率(Accuracy)进行对比,同时考查情感词典扩充及本文模型的有效性。

传统情感词典分类方法的实验分为 3 组:第 1 组实验 Dic,采用没有扩充的情感词典;第 2 组实验 Dic+SO-PMI,采用传统 SO-PMI 算法扩充领域情感词典和基础情感词典;第 3 组实验 Dic+word2vec,采用改进的 word2vec 与 SO-PMI 相结合的方法扩充领域情感词典和基础情感词典。

传统机器学习方法的实验根据分类器不同分为 4 组:第 1 组 LR,采用逻辑回归算法;第 2 组 RF,采用随机森林算法;第 3 组 NB,采用贝叶斯算法;第 4 组 SVM,采用支持向量机算法。

本文模型的实验 1 组,实验共 8 组。将设计的 8 组实验在酒店评论和京东评论上进行情感分类,各方法分类正确率如表 3 所列。

表 3 不同方法的情感分类正确率

Table 3 Correct rate of emotion classification

(单位:%)

模型	酒店评论	京东评论
Dic	79.32	76.53
Dic+SO-PMI	81.42	79.27
Dic+word2vec	81.98	80.65
LR	83.83	82.65
RF	82.50	82.33
NB	82.00	81.67
SVM	84.50	83.12
本文模型	88.30	87.40

从表 3 的结果可以看出,无论是酒店评论,还是京东评论,本文模型都取得了情感分类模型的最高正确率。本文模型在酒店评论中取得了 88.30%的正确率,相比于 Dic+word2vec 实验提高了 6.32%,相比于 SVM 提高了 3.8%;在京东评论中取得了 87.4%的正确率,相比于 Dic+word2vec 实验提高了 6.75%,相比于 SVM 提高了 4.28%。因此,实验验证了本文模型的有效性。

在情感词典方法的 3 组实验中,本文提出的 word2vec 与 SO-PMI 结合扩充领域词典方法的正确率最高,在酒店评论中正确率为 81.98%,比 Dic+SO-PMI 实验的 81.42%高出 0.56%,比 Dic 实验的 79.32%高出 2.66%;在京东评论中正确率为 80.65%,比 Dic+SO-PMI 实验的 79.27%高出 1.38%,比 Dic 实验的 76.53%高出 4.12%。这充分考虑语义信息扩充领域词典,能提高情感词典分类方法的有效性。

在机器学习方法的 4 组实验中,支持向量机算法的效果最好,在酒店评论中的正确率为 84.50%,在京东评论中的正确率为 83.12%,都高于其他 3 种方法。

本文模型结合了情感词典方法和机器学习方法的特点,构建了一个既能自动标注数据,又能提高情感倾向模糊评论的分类正确率的模型。该模型在扩充词典时考虑了词汇的语义信息,利用情感词典的分类结果自动确定情感倾向明显的评论对机器学习分类器进行训练,无须对训练集进行标注,训练过程是自监督学习;然后利用训练好的分类器对不易区分情感倾向的评论进行分类。

结合表 3 的实验结果,本文模型和传统情感词典及机器

学习方法在不同数据集上的情感分类正确率如图 2 所示。

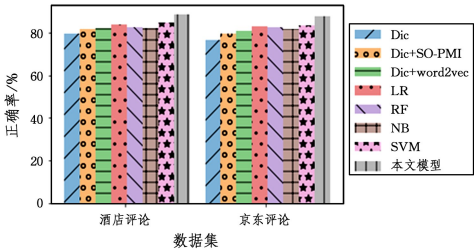


图 2 不同方法的情感分类对比结果

Fig. 2 Comparison results of emotion classification methods

为了进一步分析情感分类模型的分类结果,本文在实验过程中计算了数据集上积极文本和消极文本的精确率(Precision)、召回率(Recall)、F1。对酒店评论的分类结果如表 4 所列,对京东评论的分类如表 5 所列。

表 4 不同模型在酒店评论上的结果

Table 4 Results of different models on hotel reviews (单位: %)

模型	情感倾向	Precision	Recall	F1
Dic	积极	78.17	76.73	77.44
	消极	77.88	81.90	79.84
Dic+SO-PMI	积极	83.15	78.80	80.92
	消极	79.85	84.03	82.33
Dic+word2vec	积极	84.06	78.93	81.41
	消极	80.14	85.03	82.51
LR	积极	82.43	86.00	84.18
	消极	85.37	81.67	83.48
RF	积极	81.15	84.67	82.87
	消极	83.97	80.33	82.11
NB	积极	80.77	84.00	82.35
	消极	83.33	80.00	81.31
SVM	积极	82.86	87.00	84.88
	消极	83.90	85.02	84.10
本文模型	积极	88.38	88.20	88.29
	消极	88.22	88.40	88.31

表 5 不同模型在京东评论上的结果

Table 5 Results of different models on JD reviews (单位: %)

模型	情感倾向	Precision	Recall	F1
Dic	积极	75.70	73.21	74.43
	消极	74.06	76.49	75.26
Dic+SO-PMI	积极	78.60	76.10	77.33
	消极	76.83	79.28	78.04
Dic+word2vec	积极	81.13	78.39	79.74
	消极	79.09	81.77	80.41
LR	积极	81.82	84.00	82.90
	消极	83.56	80.00	81.74
RF	积极	81.29	84.00	82.62
	消极	83.45	80.67	82.04
NB	积极	81.58	82.67	82.12
	消极	83.45	80.67	82.04
SVM	积极	82.47	84.67	83.56
	消极	84.25	82.00	83.11
本文模型	积极	87.89	87.07	87.48
	消极	86.92	87.73	87.32

在情感词典方法中,消极评论的表现比积极评论的表现要好。酒店评论中 Dic+word2vec 实验的消极评论的 F1 为 82.51%,比积极评论的 81.41% 高出 1.1%;京东评论中 Dic+word2vec 实验的消极评论的 F1 为 80.41%,比积极评论的 79.74% 高出 0.67%。在机器学习方法中,积极评论的

表现比消极评论的表现要好。酒店评论中 SVM 实验的积极评论的 F1 为 84.88%,比消极评论的 84.10% 高出 0.78%,京东评论中 SVM 实验的积极评论的 F1 为 83.58%,比消极评论的 83.42% 高出 0.16%。

本文模型在积极评论和消极评论中取得最高的 F1,在表 4 中,本文模型在积极评论和消极评论中的 F1 分别为 88.29%和 88.31%;在表 5 中,本文模型在积极评论和消极评论中的 F1 分别为 87.48%和 87.32%,这进一步证明了本文模型的有效性。

结束语 本文根据网络评论的特点,在已有研究成果的基础上,提出了融合情感词典和机器学习的分类模型,经实验验证,该模型的分类正确率较其他方法有所提升。本文根据网络评论的领域特点,通过 word2vec 与 SO-PMI 相结合的方法扩充领域词典,依据词典和规则对网络评论进行情感分类,设置情感阈值确定机器学习分类器的训练集,自动标注训练集,通过自监督学习训练机器学习分类器,以进一步改进词典的分类结果。今后的研究工作还需要进一步完善领域词典的扩充,使用更全面的语言规则来提高自监督训练语料的质量,以获得更好的分类效果。

参 考 文 献

[1] HONG W,LI M. A Summary of Research on Text Sentiment Analysis Methods[J]. Computer Engineering & Science, 2019, 41(4):750-757.

[2] QIU L,ZHANG W, HU C, et al. Selc: a self-supervised model for sentiment classification[C]// Proceedings of the 18th ACM Conference on Information and Knowledge Management. 2009: 929-936.

[3] HATZIVASSILOGLOU V,MCKEOWN K R. Predicting the semantic orientation of adjectives[C]// Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics, 1997:174-181.

[4] WIEBE J. Learning subjective adjectives from corpora[C]// Proceedings of the 17th National Conference on Artificial Intelligence. Menlo Park, CA: AAAI Press, 2000:735-741.

[5] TURNEY P D,LITTMAN M L. Measuring praise and criticism: Inference of semantic orientation from association[J]. ACM Transactions on Information Systems (TOIS), 2003, 21(4):315-346.

[6] LI S S,LI Y W,HUANG J R, et al. Construction method of Chinese sentiment dictionary based on bilingual information and label propagation algorithm[J]. Journal of Chinese Information Processing, 2013, 27(6):75-81.

[7] WANG Z T,YU Z W, GUO B, et al. Sentiment Analysis of Chinese Weibo Based on Dictionary and Rule Set[J]. Computer Engineering and Applications, 2015, 51(8):218-225.

[8] FAN Z, GUO Y, ZHANG Z H, et al. Sentiment analysis of movie reviews based on dictionaries and weakly annotated information[J]. Journal of Computer Applications, 2018, 38(11):3084-3088.