

引用格式:郑诚,潮旭,章金平. 基于规则的情感本体和词向量的中文情感分类[J]. 微电子学与计算机,2019,36(6):50-54.
[ZHENG C, CHAO X, ZHANG J P. Chinese sentiment classification of emotional ontology and word vector based on rules
[J]. Microelectronics & Computer,2019,36(6):50-54.]

基于规则的情感本体和词向量的中文情感分类

郑 诚^{1,2}, 潮 旭^{1,2}, 章金平^{1,2}

(1 安徽大学 计算智能与信号处理实验室,安徽 合肥 230601;2 安徽大学 计算机学院,安徽 合肥 230601)

摘 要: 随着电子商务的快速发展,评论文本的情感倾向研究引起了广大学者的关注. 为了充分利用短文本中的情感本体和语义信息,提出结合句法规则、情感本体和词向量的中文情感分类方法. 首先利用 Word2vec 训练词向量,结合句法规则生成短文本向量;再根据情感特征分布,创建领域自适应情感词典,结合句法规则,得到短文本情感值,从而构建词向量和情感值相结合的情感模型 VWEO(Vector with Emotional Ontology). 在酒店评论数据集中,与已有方法相比,所提方法在准确率、召回率、F1 值均有明显提升.

关键词: 句法规则; Word2vec; 情感本体; 情感词典

中图分类号: TP391

文献标识码: A

文章编号: 1000-7180(2019)06-0050-05

Chinese sentiment classification of emotional ontology and word vector based on rules

ZHENG Cheng^{1,2}, CHAO Xu^{1,2}, ZHANG Jin-ping^{1,2}

(1 Key Laboratory of ICSP Ministry of Education. Anhui University, Hefei 230601, China;

2 School of Computer Science and Technology. Anhui University, Hefei 230601, China)

Abstract: With the rapid development of e-commerce, the research on the emotional tendency of comment texts has attracted the attention of scholars. In order to make full use of the emotional ontology and semantic information in short texts, a Chinese sentiment classification method combining syntactic rules, emotional ontology and word vector is proposed. Word2vec is firstly used to train word vectors and generate short text vectors in conjunction with syntactic rules; Then, according to the distribution of emotional features, a domain adaptive sentiment dictionary is created, and syntactic rules are combined to obtain short text sentiment values, thereby constructing an emotional model called VWEO(Vector with Emotional Ontology) combining word vectors and emotional values. In the hotel review data set, compared with the existing methods, the proposed method has significantly improved accuracy, recall rate and F1 value.

Key words: syntactic rules; word2vec; emotional ontology; emotional dictionary

1 引言

近年来,随着社交网络的快速发展,互联网上产生大量用户对于产品、人物、事件的评论数据,而情感分类的任务就是识别出用户在评论文本中流露出

的情感信息,将其分为代表肯定和赞扬的正类以及代表否定和批判的负类. 常见的情感分类方法一般分为基于词典的规则方法和统计机器学习方法,也不断有研究者将两种方法进行结合与探索.

基于词典的方法是根据情感词典中的情感信息

计算评论文本的情感倾向性^[1-2].

统计机器学习方法需要有标注的数据来建立分类模型来预测文本情感倾向. 由于其出色的分类效果,得到越来越多研究者的青睐,成为了情感分类主流方法^[3-4]. 以上方法大部分是仅限于挖掘出句子中词与词之间的词汇特征和句法特征,并未考虑语义特征.

自 2013 年 Mikolov 等^[5]提出 word2vec 这一概念后,word2vec 提取词语语义关系的方法也被初步应用到情感分类任务上. Bai 等^[6]通过语义定位点相似距离(SO-SD)模型,利用 Word2vec 工具创建出一种新的模型用以建立情感词典,并使用该情感词典来获取微博消息的情感倾向 Zhang 等^[7]使用 word2vec 对相似的特征进行聚类,再分别使用基于词典和基于词性的特征选择方法与 SVM^{perf}相结合进行情感分类.

本文为了提取语义关系,使用 word2vec 工具获取词向量,并利用评论数据自动提取情感种子词,从而创建领域自适应情感词典,与规则方法相结合,构成短文本词向量和情感值相融合情感模型(VWEO),考虑了情感本体和语义信息,从而达到提高情感分类效果的目的.

2 word2vec 训练词向量

Word2vec 是 Google 在 2013 年推出的一款面向大众的开源的用于训练词向量的工具,它提供了一种使用分布式向量对文本进行表示的方法. 它是一种神经网络概率语言模型,其主要目的是将单词转换为包含语义的词向量. 它有两种训练模式,分别是通过上下文预测给定词的 CBOW(Continuous Bag of Words)模型以及根据当前词语来预测上下文的 Skip-gram 模型. 本文使用默认的 Skip-gram 模型对词向量进行训练.

同一评价对象,不同的人对其有不同的称呼,为降低评价对象特征词对分类结果的影响,本文利用 Word2vec 计算余弦相似度,对于最常见的评价对象特征进行聚类,如表 1 所示,聚类后用统一评价对象进行表示.

表 1 相似评价对象特征聚类

评价对象	相似评价对象
房间	房屋 房子 酒店房间
设施	设备 配置 硬件
价格	房价 价钱 价位
交通	出行 酒店交通 酒店出行
总体	整体 综合 酒店总体

3 基于句法规则的情感本体与词向量结合的方法

3.1 特殊句式的挖掘

汉语语句通常分为简单句式和复杂句式. 本文主要针对复杂句式的转折句式进行研究.

转折句一般指它所包含的前后分句之间存在转折关系,句中包含转折连词,后一部分往往是作者的真实想法. 转折连词分为“虽然”、“尽管”这类让步词以及“但是”、“然而”这类词. 一般可以将转折句分为三类:第一类是完整的转折句式,包含两种转折连词;第二类缺少第一种转折连词;第三类缺少第二种转折连词. 如表 2 所示:

表 2 转折句式举例

类型	样本
第一类	尽管酒店服务不错,但我不是很喜欢
第二类	酒店环境还行,但整体有问题
第三类	这酒店真的符合我的品位,尽管设施有点旧

对转折句式进行情感分析,要充分考虑它的结构特点易知,最终的情感倾向通常在第二类转折连词之后,而与第一类转折连词相反,所以第一类转折连词起到的作用与否定词相似. 因此提出如下规则:

- (1)第一类转折特征词取特征值-1.0;
- (2)第二类转折特征词取特征值 1.0.

3.2 情感资源构建

特征词典:通过对比 DF(Document Frequency)、IG(Information Gain)、MI(Mutual Information)以及 X²等(CHI)特征选择方法,选择出分类效果最好的特征.

程度副词词典:由 Hownet 程度副词和特征词典相交的部分所构成. 根据程度等级分别赋予特征值 2.0、1.5、1.2、0.8、0.5、-1.0. 如表 3 所示:

表 3 程度副词特征值

程度词	特征值
最、充分、极度	2.0
很、格外、太	1.5
更、足、较	1.2
稍、有点、挺	0.8
相对、不怎么、微	0.5
过于、过分	-1.0

否定词词典:本文收集了一些常用否定词,并赋予特征值-1.0. 如表 4 所示:

表 4 否定词特征值

否定词	特征值
不、非、从未、决不	-1.0

3.3 领域自适应情感词典的构建

目前,大部分学者基于已有的情感词典来进行评论文本的情感倾向研究,如 HowNet 情感词典、清华大学的情感词典.而针对特定的领域,这些情感词典往往存在着一些不足,如:领域词、多义词等,已有的情感词典并不能对特定领域表达出正确的情感倾向,所以本文构建了一个领域自适应情感词典(只考虑形容词).这里构建领域自适应情感词典主要涉及两步:

(1) 统计每个词 w 分别在正类和负类中的相对频率.如公式(1)所示:

$$\begin{cases} p_+(w) = \frac{num(w, +)}{Sum_+} \\ p_-(w) = \frac{num(w, -)}{Sum_-} \end{cases} \quad (1)$$

式中, $num(w, +)$ 是正类中目标词 w 的数量; Sum_+ 是正类中单词总数; $num(w, -)$ 是负类中目标词 w 的数量; Sum_- 是负类中单词总数.

(2) 判断词 w 的情感倾向值.如公式(2)所示.

$$D(w) = p_+(w) - p_-(w) \quad (2)$$

式中, $D(w)$ 为词 w 在正类和负类中频率的差值.如果 $D(w)$ 的值是正数,则单词 w 是正面情感词;如果 $D(w)$ 的值是负数,则单词 w 是负面情感词^[3].

通过上述方法得到的各情感词情感倾向值过于依赖情感词数量,因此不作为各情感词最终权重.选取情感权重最大的 10 个正面情感词以及情感权重最小的 10 个负面情感词作为情感种子词,利用词向量计算每个情感词分别与正面情感种子词以及负面情感种子词的欧式距离,如式(3)所示:

$$\begin{cases} dist_+ = \sum_{i=1}^{10} \|W - S_i\| \\ dist_- = \sum_{j=1}^{10} \|W - S_j\| \end{cases} \quad (3)$$

式中, S_i 是正面情感种子词词向量; S_j 是负面情感种子词词向量; W 是目标词词向量; $dist_+$ 是目标词与正面情感种子词总欧式距离; $dist_-$ 是目标词与负面情感种子词总欧式距离.

对于单词 w ,通过比较 $dist_+$ 和 $dist_-$ 来确定正面情感词和负面情感词,计算出差值 $Q(w)$,如式(4)所示,并将该差值赋给单词 w ,它代表单词 w 的最终权重.

$$Q(w) = dist_- - dist_+ \quad (4)$$

如果 $Q(w)$ 的值是正数,则单词 w 是正面情感词;如果 $Q(w)$ 得值是负数,则单词 w 是负面情感词,同时 $Q(w)$ 代表单词 w 的情感强度值.

3.4 特征提取规则

特征的提取,是情感分类任务中的关键.本文考虑了语义规则,根据 Word2vec 训练得到词向量,然而词向量并不包含情感倾向,语义相近的词的词向量并不一定情感倾向相似,因此提出如下规则:

(1) 程度副词特征:当情感词前面出现程度副词时,以程度副词的特征值乘以该情感词的词向量.

(2) 否定词特征:统计出现在情感词之前的否定词次数,将否定词特征值相乘,若结果为 1.0,则以 1.0 乘以该情感词词向量;若结果为 -1.0,则将该情感词之前修饰它的程度副词,否定词以及该词本身的词向量不计入文本总词向量.

(3) 转折句式特征:当情感词前面出现第一种转折连词时,将该转折连词的特征值 -1.0 与否定词特征值相乘,结果若为 1.0,则以 1.0 乘以该情感词词向量;结果若为 -1.0,则将该情感词之前修饰它的程度副词,否定词以及该词本身从短文本中删除.当情感词前面出现第二种转折连词时,按 1,2 两种规则处理.

(4) 情感词特征:它根据程度副词、否定词、转折连词以及情感词本身的词向量综合计算而来,它的具体计算公式如公式(5)所示:

$$\begin{aligned} senti(word) = & f_1(x_1) * \prod_{i=1}^n f_2(x_2) \\ & * \prod_{j=1}^n f_3(x_3) * \text{vector}(word) \end{aligned} \quad (5)$$

式中, $f_1(x_1)$ 是转折连词的特征值; $f_2(x_2)$ 是否定词的特征值; $f_3(x_3)$ 是程度副词的特征值.若 $f_1(x_1)$ 、 $f_2(x_2)$ 、 $f_3(x_3)$ 的乘积小于 0,则将该情感词之前修饰它的程度副词,否定词以及该词本身的词向量不计入文本总词向量.

(5) 评论文本表示:评论文本特征向量,由各特征词的词向量按维度相加后平均表示,如公式(6)所示:

$$senti(doc) = \frac{1}{n} \sum_{i=1}^n \text{vector}(word_i) \quad (6)$$

式中, $word_i$ 是特征词典中的特征.

3.5 情感值计算

每条评论的情感值由其情感词的数量以及情感词的情感强度来决定,如公式(7)所示:

$$T = \sum_{i=1}^n f(x_1) * f(x_2) * f(x_3) * Q_i \tag{7}$$

式中, Q_i 表示第 i 个情感词的情感强度; $f(x_1)$ 是转折连词的特征值; $f(x_2)$ 是否定词的特征值; $f(x_3)$ 是程度副词的特征值; T 表示该评论的总情感值.

3.6 算法描述

根据前几节的思想,将本文提出方法中核心算法描述如下.

基于规则词向量的分类算法

输入:分好词的评论文本

输出:代表该评论的特征向量

开始

步骤 1:计算各情感词经程度副词、否定词以及转折连词修饰后的词向量之和,并计算评论文本情感总值

步骤 2:计算文本除去形容词后与特征词典相交的词的特征向量之和

步骤 3:将步骤 1 与步骤 2 的词向量之和相加

步骤 4:将总词向量与该评论情感总值拼接后作为该评论的特征向量

结束

4 实验结果及分析

4.1 数据集与数据的预处理

本文数据集来自谭松波老师公布的公用酒店数据集,以及从携程网爬取的共计 12 万条酒店评论数据.对爬取数据集进行去重、去除过短的文本等处理,剩余的 10 万余条评论用于训练词向量.再对谭松波老师公布的公用酒店数据集的 10000 条数据进行去除重复、不完整以及篇幅过短的处理后,选取了正负评论各 2000 条作为测试数据.

4.2 评价指标

本文的评价标准是准确率(Precision)、召回率(Recall)和 F 值(F-measure).准确率是测试集中正确分类的评论文本数占分类结果中所有标记为该类文本数量的比例;召回率则为测试集中正确分类的评论文本数占测试集中人工标注的所有属于该类文本数量的比例;F 值是两者的调和平均.它们的具体计算公式如公式(8)所示:

$$P = \frac{class_correct}{class_proposed}$$
$$R = \frac{class_correct}{class_labeled} \tag{8}$$
$$F-measure = \frac{2 \times P \times R}{P + R}$$

4.3 实验结果与分析

为了选出最佳的分类特征,本文在特征数为 500

的前提下,分别利用 DF(Document Frequency)、IG(Information Gain)、MI(Mutual Information)以及 X^2 (CHI)等特征选择方法选择特征,结合布尔模型以及 SVM 算法,得出的结果如表 5 所示:

表 5 不同特征选择方法下的情感分类结果

特征	准确率	召回率	F 值
DF	0.845	0.843	0.844
IG	0.857	0.858	0.858
MI	0.585	0.585	0.585
X^2	0.862	0.860	0.861

实验结果显示 X^2 特征选择方法最适合情感分类,于是本文在该特征选择方法下继续实验,选择不同的特征数量结合词向量进行实验,实验结果如表 6 所示:

表 6 不同特征数量下的情感分类结果

特征	准确率	召回率	F 值
500	0.874	0.874	0.874
1000	0.880	0.882	0.881
1500	0.884	0.884	0.884
2000	0.884	0.885	0.885
2500	0.888	0.889	0.888
3000	0.889	0.889	0.889
3500	0.890	0.890	0.890
4000	0.889	0.888	0.888
4500	0.887	0.887	0.887

根据实验结果可知,在特征数量为 3500 的情况下,情感分类效果最好,因此确定特征数量.

本文将基于词典的方法(DICT),考虑特征词向量的方法(FV)^[7],以及本文的基于句法规则的词向量的方法(SRFV)和 VWEO 方法进行了比较,实验结果如表 7 所示:

表 7 分类结果比较

分类方法	准确率	召回率	F 值
DICT	0.754	0.752	0.753
FV	0.890	0.890	0.890
SRFV	0.894	0.895	0.894
VWEO	0.900	0.899	0.900

由以上实验结果可知,在考虑句法规则的情况下,情感分类结果相对于只考虑特征词向量的方法更好,当加入情感本体后,情感分类效果最好,验证了本文方法的可行性,句法规则的词向量以及情感本体的结合对分类效果起到正面影响.

5 结束语

本文针对情感分类问题,提出结合句法规则、词向量以及情感本体的方法,同时考虑句法特征和语义特征对数据进行分类,由实验数据表明,本文的VWEO方法在情感分类准确率方面相比于传统的分类方法有着一定的优越性.在以后的工作中,我们打算对被否定的情感词进行更高效的利用,并对更多的句式进行分析,而不仅仅只是考虑转折句式,进一步的使情感分类效果得到提升.

参考文献:

[1] TURNEY P. Thumbs Up or Thumbs Down Semantic Orientation Applied to Unsupervised Classification of Reviews[C]. Proceedings of Annual Meeting of the Association for Computational Linguistics, Philadelphia,USA, 2002:417-424.

[2] ALMATARNEH S, GAMALLO P. Automatic Construction of Domain-Specific Sentiment Lexicons for Polarity Classification[C]// International Conference on Practical Applications of Agents and Multi-Agent Systems, Avila,Spain, 2018.

[3] LIU K L, LI W J, GUO M. Emoticon Smoothed Language Models for Twitter Sentiment Analysis[C]//

AAAI Conference on Artificial Intelligence, Toronto, Canada, 2012.

[4] 姜杰, 夏睿. 机器学习与语义规则融合的微博情感分类方法[C]// 自然语言处理与中文计算会议. Kunming,China, 2016.

[5] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient Estimation of Word Representations in Vector Space[J]. Computer Science, 2013, 5(4):243-254.

[6] XUE B, FU C, ZHAN S. A Study on Sentiment Computing and Classification of Sina Weibo with Word2vec[C]// IEEE International Congress on Big Data, Washington,USA, IEEE, 2014:358-363.

[7] ZHANG D, XU H, SU Z, et al. Chinese comments sentiment classification based on word2vec and SVM perf[J]. Expert Systems with Applications, 2015, 42(4):1857-1863.

作者简介:

郑 诚 男,(1964-),博士,副教授.研究方向为数据挖掘,智能语义信息检索.

潮 旭(通讯作者) 男,(1994-),硕士研究生.研究方向为数据挖掘,情感分类,大数据. E-mail:1105741346@qq.com.

章金平 女,(1991-),硕士研究生.研究方向为数据挖掘,推荐系统,大数据.

(上接第 49 页)

[6] 赵小国,刘丁,景坤雷. 融合改进蚁狮算法和 T-S 模糊模型的噪声非线性系统辨识[J/OL]. 控制与决策:1-7 [2018-06-11]. <https://doi.org/10.13195/j.kzyjc.2017.1282>.

[7] 景坤雷,赵小国,张新雨,等. 具有 Levy 变异和精英自适应竞争机制的蚁狮优化算法[J]. 智能系统学报, 2018,13(2):236-242.

[8] 杜晓昕,张剑飞,郭媛,等. 基于柯西-高斯动态消减变异的果蝇优化算法研究[J]. 计算机工程与科学,2016, 38(6):1171-1176.

[9] 郭振洲,王平,马云峰,等. 基于自适应权重和柯西变异的鲸鱼优化算法[J]. 微电子学与计算机,2017,34(9): 20-25.

[10] 韩旭明,邱兵,刘翹铭,等. 基于柯西变异的果蝇优化算法[J]. 微电子学与计算机,2017,34(11):26-30.

作者简介:

于建芳 女,(1992-),硕士研究生.研究方向为群智能算法、智能计算、工程优化、项目调度.

刘 升(通讯作者) 男,(1966-),博士,教授.研究方向为人工智能、智能计算. E-mail:m18817568572_1@163.com.

韩斐斐 女,(1993-),硕士研究生.研究方向为群智能算法、智能计算.

肖子雅 女,(1994-),硕士研究生.研究方向为智能算法、项目调度与优化.