

扩展语义相似情感词的文本情感分类方法

罗森林，毛焱颖，潘丽敏，陈倩柔，魏超

(北京理工大学 信息系统及安全对抗实验中心,北京 100081)

摘要：针对文本情感分类中情感语义特征利用不足、特征降维效果欠佳等影响分类效果的问题,提出了一种通过扩展语义相似的情感词以及引入词语间统计特征的高精度网络评论情感分类方法。该方法利用神经网络 Skip-gram 模型生成词嵌入,通过词嵌入相似性度量将语义相似的词语扩展为情感特征;再利用词语间的统计特征进行特征降维;通过多个弱分器加权构建 Adaboost 分类模型实现网络评论情感分类。基于酒店评论和手机评论公开测试集进行实验,结果表明其情感分类的正确率分别达到 90.96% 和 93.67%。方法扩展语义相似情感词有利于丰富文本情感语义特征,引入词语间的统计特征有更好的特征降维效果,可以进一步提升文本情感分类的效果。

关键词：词嵌入; Adaboost 分类模型; 特征选择; 中文评论; 情感分类

中图分类号: TP 391 文献标志码: A 文章编号: 1001-0645(2018)11-1156-08

DOI: 10.15918/j.tbit.1001-0645.2018.11.009

A Method of Text Sentiment Classification by Extending Semantic Similar Sentiment Words

LUO Sen-lin, MAO Yan-ying, PAN Li-min, CHEN Qian-rou, WEI Cao

(Information System and Security & Countermeasures Experimental Center, Beijing Institute of Technology, Beijing 100081, China)

Abstract: To solve the effect problem of sentiment classification due to the insufficient use of emotional semantic features and unpromising dimension reduction effects, a novel high-precision sentiment classification method was proposed in this paper for online comments by extending semantic similar emotional words and employing the statistical features between words. Firstly, a neural network skip-gram model was employed to generate word embedding and extend the semantic similar words to emotional feature by the measure of embedding word similarity. Then the feature dimension was reduced by employing the statistical features between words. At last, sentiment classification of online comments was carried out by the Adaboost classification model which was constructed by weighting multiple weak classifiers. Experiment results on hotel reviews and mobile comments show that, the accuracy of sentiment classification with new method can reach 90.96% and 93.67% respectively. Expanding semantic similarity emotion words is helpful to enrich the semantic features of emotion. Employing statistical features between words has better feature reduction effect. Both two procedures effectively improve the performance of text sentiment classification.

Key words: word embedding; Adaboost classification model; feature selection; Chinese comment; sentiment classification

收稿日期: 2017-06-23

基金项目: 北京理工大学基础研究基金资助项目(20160542013); 国家“二四二”计划项目(2017A149)

作者简介: 罗森林(1968—),男,教授,博士生导师,E-mail:luosenlin@bit.edu.cn.

通信作者: 毛焱颖(1993—),女,硕士,E-mail:1023017632@qq.com.

随着电商技术的迅速发展,针对商品和服务的评论性文本大量出现。这些网络评论包含丰富的情感信息,如喜、怒、哀、乐等^[1]。为了更好地抽取这些评论中的情感信息,提升电商服务的质量,情感分析应运而生。情感分析(sentiment analysis),又被称为观点挖掘(opinion mining),是通过特定的方法分析、归纳和推理出作者想要表达的情感、观点和立场,进而可以通过情感信息将其划分为不同的情感倾向类^[2]。

常见的文本情感分析方法主要有两类:基于情感词典的方法和基于机器学习的方法^[1]。基于词典的方法主要通过情感词典和制定一系列的相关规则,对文本进行依存句法分析,然后计算情感值,最后根据情感值判定文本的情感倾向^[3]。基于机器学习的方法是将情感分析问题转化为一个分类问题,将目标情感分成两类:正、负,对训练文本进行人工标注,然后进行情感分类。基于词典的方法在构建情感词典时需要大量人力,且建立和维护规则集的过程比较费时费力,而基于机器学习的有监督分类方法根据标注的样本集进行分类器的训练和测试,分类精确度高。目前,在国内外研究中,基于机器学

习的有监督文本情感分类方法主要有:K 近邻方法(KNN)^[4]、朴素贝叶斯方法(NB)^[5]、最大熵方法(ME)^[6]、支持向量机(SVM)^[7]、Adaboost^[8]等。

进行有监督情感分类时,用户的观点错综复杂,常常发表混合观点评论,既肯定某方面,同时又在批评另外的方面,观点评论噪音多、信息量大、情感表达复杂。对评论进行情感分类时提取情感特征困难,容易导致情感语义特征利用不足、特征降维效果欠佳,影响情感分类效果。为了处理这些问题,本文提出一种基于词嵌入相似性度量扩展语义相似情感词的中文评论情感分类方法,该方法能丰富文本情感语义特征、提升特征降维效果,进一步提高情感分类的效果。

1 中文评论情感分析方法

1.1 原理框架

本方法原理图如图 1 所示。对中文网络评论文本,基于 Word2vec 的 Skip-gram 模型生成词嵌入,利用词嵌入的相似性度量选取语义相似的词语,扩展语义相似情感词为情感特征,利用词语间统计特征进行特征降维,引入 Adaboost 分类模型进行情感倾向性分类。

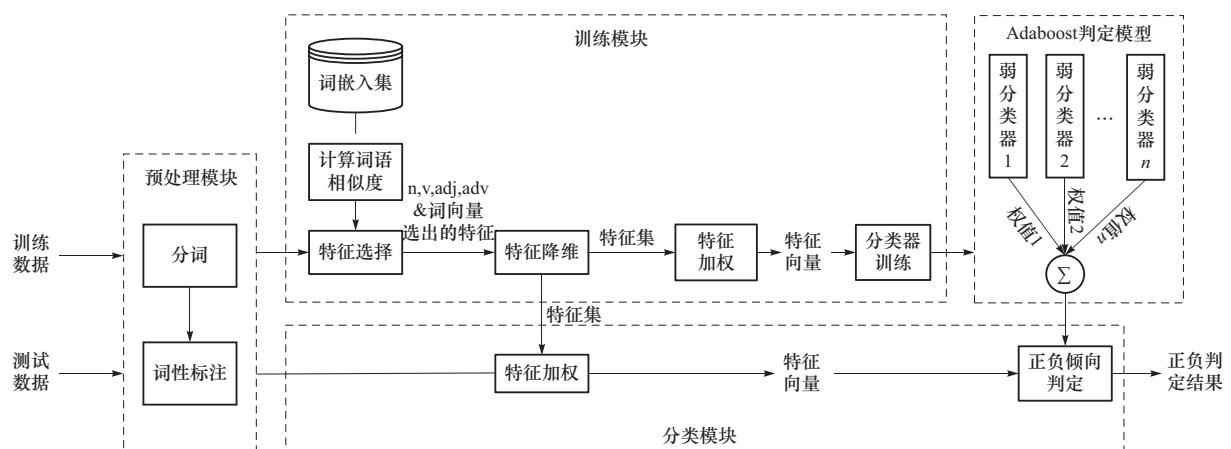


图 1 扩展语义相似情感词的情感分类方法原理图

Fig. 1 Theory of sentiment classification method based on extending semantic similar emotion words

1.2 计算词语相似度

词嵌入是自然语言处理的表征学习技术和语言模型的总称,它将一个单词或词组映射到低维空间的实数向量。词嵌入可以由经典的神经网络语言模型(NNLM)^[9]在训练中产生,NNLM 去掉隐藏层改进后可构建 Word2vec 模型。Word2vec 是一款用于计算词嵌入的工具,由 Google 在 2013 年开源,它通过优化后的训练模型,根据给定的语料库,将一个词

语表达成向量形式。Word2vec 使用的词嵌入表示方式是分布式表示,采用一个 3 层模型:输入层、投影层和输出层。这个 3 层神经网络本身是对语言模型进行建模,但 Word2vec 的真正目标是在建模同时获得一种单词在向量空间上的表示。Word2vec 包含两种模型,分别是 CBOW 模型和 Skip-gram 模型,两种模型都是由 NNLM 改进而来,由图 2 可见,其中 $w(t)$ 代表当前词语位于句子的位置 t 。

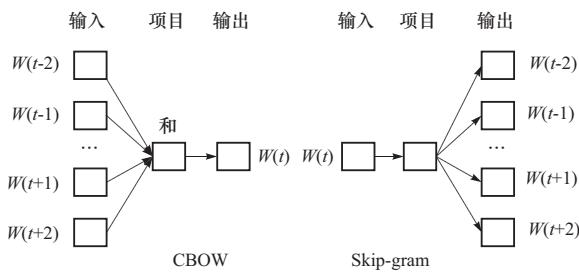


图2 CBOW 和 Skip-gram 模型

Fig. 2 CBOW model and Skip-gram model

CBOW 是根据上下文来预测当前词语的概率, 去掉了 NNLM 的非线性部分. Skip-gram 则是根据当前词语来预测上下文的概率. CBOW 模型的目标函数公式如式(1).

$$L = \sum_{w \in C} \lg p(w | \text{Context}(w)). \quad (1)$$

式中: $\text{Context}(w)$ 表示的是该词的上下文; C 表示语料库; L 为目标函数. 通过梯度下降法优化目标函数, 通过求目标函数的最大值可以得到最优模型参数. Skip-gram 模型的目标函数公式如式(2).

$$L = \sum_{w \in C} \lg p(\text{Context}(w) | w). \quad (2)$$

式中: $\text{Context}(w)$ 表示的是该词的上下文; C 表示语料库; L 为目标函数. 本文 Word2vec 使用 Skip-gram 模型.

Word2vec 能将词表征为实数值向量, 通过给定的语料进行训练, 把对文本内容的处理简化为向量空间中的向量运算, 利用了神经网络的思想. 通过向量空间中的相似度表示词语语义上的相似度. 两个词向量间语义相似度的取值范围为 0~1 之间, 值越大代表两个词的关联度越高. 词嵌入 $N = (w_1, w_2, \dots, w_n)$, 词嵌入 N_1 和 N_2 的相似度计算公式如式(3).

$$\begin{aligned} \text{sim}(N_1, N_2) &= \cos \theta = \\ &= \frac{\sum_{k=1}^n w_k(N_1) w_k(N_2)}{\sqrt{\left(\sum_{k=1}^n w_k^2(N_1)\right) \left(\sum_{k=1}^n w_k^2(N_2)\right)}}. \end{aligned} \quad (3)$$

1.3 特征选择

文本特征选择中选择的特征不仅要正确地反映文本的情感线索, 还要清楚地区分一个文档和其他文档, 可以是词、词的组合等. 对每一个评论文本进行特征选择时, 选择 $N, V, \text{Adj}, \text{Adv}$ 的组合 NVAA 作为文本特征. 但是 NVAA 包含的作者情感有限, 如果特征过少容易导致情感语义特征利用不足, 本

文首先训练 Word2vec 生成词嵌入, 通过词嵌入计算评论中其他词语与 NVAA 中每个特征的语义相似度, 根据语义相似度选择相似的特征添加到 NVAA 特征中. 特征选择过程中, 由于否定词会转换语义, 如果有否定词“不”和“不是”, 这两词和接下来邻近的形容词、副词、动词构成一个词组作为特征, 如特征“不喜欢”, “不舒服”等. 本文特征选择过程如下.

输入: 分词并词性标注的评论文本集

输出: 特征选择后的评论文本集

过程: for 评论文本集

 for 每条评论文本

 for 每个词语

 词语若是否定词“不”或“不是”, 则和接下来的词语合并成一个短语, 选择该短语和词性为 $N, V, \text{Adj}, \text{Adv}$ 的词语添加进特征集

 end for

 通过词嵌入计算评论中其他词语与词性为 $N, V, \text{Adj}, \text{Adv}$ 的词语的语义相似度, 选出相似的词语添加进特征集

 end for

end for

1.4 特征降维

一般基于统计的特征降维方法包括: DF、IG、MI^[10], 利用该常用特征降维方法进行特征降维效果欠佳. 为提升特征降维效果, 引入词频, 利用词语间的统计特征进行特征降维, 本文提出 DF-TF、IG-TF、MI-DF.

由于 DF 考虑特征在多少个文档中出现, 没有考虑在一个文档中出现的次数. 针对一个特征, 它在多个文档中出现, 且在每个文档中出现多次, 具有较大的区分度. 在一个文档中出现多次的特征在同类别的其他文档中也会出现多次. 在进行特征降维时, 选取高词频特征词可以提高文本分类的性能^[11]. 所以加入词频的影响因素, 改进后的公式如式(4).

$$\text{DF-TF} = \text{DF} \times \frac{\text{lb TF}}{1 + \text{lb arg(TF)}}. \quad (4)$$

式(4)中 arg(TF) 是特征项在每篇文档中 TF 值的均值.

IG 考虑了特征项不存在时的信息量, 加入词频后可以减少不必要的干扰. 且 IG 衡量的是每个特征对整个分类系统的贡献度, 不会具体到某类文档

中. IG-TF 引入词频因子,选出的特征集中可以分布到某个类别,且可以在该类别合理平均分布,改进后的公式如式(5).

$$\begin{aligned} \text{IG-TF}(t) = & - \sum_{i=1}^n p(c_i) \ln p(c_i) + \\ & p(\bar{t}) \sum_{i=1}^n p(c_i | \bar{t}) \ln p(c_i | \bar{t}) + \alpha \times \\ & p(t_1) \sum_{i=1}^n p(c_i | t) \ln p(c_i | t). \end{aligned} \quad (5)$$

式中: $p(c_i)$ 表示类别出现的概率,如果文档集是均等的,类别为 2,那么 $p(c_i)$ 值就为 $1/2$; $p(c_i | t)$ 表示特征 t 出现的情况下,文档为类别 C_i 的概率,即为含有特征 t 同时属于类别 C_i 的文档数除以特征 t 的 DF 值; $p(t_1)$ 为特征 t 词频为 1 时候的概率, α 为经验参数.

MI 测量的是特征项和类别之间的统计关系,更倾向于选择低频词作为情感特征;DF 值低于某个阈值的词不含有或含有较少的类别信息. MI-DF 引入文档频率,可以去掉部分低频词,从而覆盖更多有效特征,改进后如式(6).

$$\begin{aligned} \text{MI-DF}(t, C_i) = & \text{DF} \times \lg \frac{p(t, C_i) + 1}{p(t)p(C_i) + 1} = \\ & \text{DF} \times \lg \frac{AN + 1}{(A+C)(A+B) + 1}. \end{aligned} \quad (6)$$

式中: t 表示特征项; C_i 表示类别; A 表示类别 C_i 包含特征项 t 的文档数量; B 表示不属于类别 C_i ,但是包含特征项 t 的文档数量; C 表示类别 C_i 中不包含特征项 t 的文档数量; D 表示不属于类别 C_i ,同时不包含特征项 t 的文档数量. 类别 C_i 的文档数量为 $M=A+C$. 训练集中所有文档的数量为 $N=A+B+C+D$.

1.5 特征加权

大多数情况下,富有情感色彩的特征出现的次数并不重要,重要的是它是否出现,以及在哪个类别中出现^[12]. 本文中使用布尔权重进行特征权重计算. 每个特征的权重值为 1 或 0,1 表示这个特征在这条评论语句中,0 表示这个特征没有在这条评论语句中.

在特征降维和特征加权之后,每一条中文评论被表示成一个一维向量,向量维度为特征降维后的维度,其中的词又称为特征,每个特征对应的权重为布尔权重. 中文评论文本集就被表示为一个单词-文档的二维矩阵,即为向量空间模型,把该二维矩阵作为分类器的输入.

1.6 分类器训练

Adaboost 算法^[8]通过改变数据权重分布实现,对训练样本进行 T 轮训练,根据样本是否被正确分类而赋予不同权值,被正确分类的样本赋予较小权值,被错误分类的样本赋予较大权值,并将更新权值后的新数据集进行下一次模型训练,样本权重同时会影响该轮弱分类器的权重. T 轮训练后,得到 T 个弱分类器,将弱分类器加权组合成强分类器,权值表示对应弱分类器的性能,权值越大表示分类器性能越好. Adaboost 算法的具体描述如下.

假设 X 表示样本空间,有 Y 表示样本类别标识. 对于二值分类问题, $Y \in \{-1, +1\}$.

训练集: $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$, 其中 $x_i \in X, y_i \in Y$.

① 初始化 N 个样本的权重 $D_0(x_i, y_i) = 1/N$, $i=1, 2, \dots, N$.

② 循环训练 T 次,对于 $t=1, 2, \dots, T$,有:

1) 对于每个特征 j ,训练一个弱分类器 h_j . 对于样本权重 $D_t(x_i, y_i)$,计算错误率 $\epsilon_j = \sum_{i: y_i \neq h_j} D_t(x_i, y_i)$,

本文采用决策树作为弱分类器;

2) 选择错误率最小的弱分类器 h_t 作为本轮训练的最终弱分类器;

3) 如果 $\epsilon_t \geq 0.5$,停止;

4) 如果 $\epsilon_t \leq 0.5$,计算弱分类器权重系数 $\alpha_t = \frac{1}{2} \ln \left(\frac{1-\epsilon_t}{\epsilon_t} \right)$;

5) 更新样本权重

$$D_{t+1}(x_i, y_i) = \frac{D_t(x_i, y_i)}{Z_t} \times \begin{cases} e^{-\alpha_t}, & h_t(x_i) = y_i \rightarrow \frac{D_t(x_i, y_i)}{Z_t} e^{-\alpha_t y_i h_t(x_i)}, \\ e^{\alpha_t}, & h_t(x_i) \neq y_i \end{cases}$$

其中, Z_t 为归一化常数,以保证 $D_t(x_i, y_i)$ 为完全概率分布, $Z_t = \sum_{i=1}^T D_t(x_i, y_i) e^{-\alpha_t y_i h_t(x_i)}$;

③ 迭代 T 次后,由 T 个弱分类器加权组合构成最终的强分类器

$$H(x) = \text{sgn} \left(\sum_{t=1}^T \alpha_t h_t(x) \right).$$

2 实验分析

2.1 实验数据

实验数据包括中文手机评论语料和中文酒店评论语料,中文手机评论语料是京东上的手机评论语

料,中文酒店评论语料是中科院谭松波博士整理的酒店评论语料。首先经过人工筛选去除了重复的评论再对每条评论的情感倾向(正向或负向)进行人工标注。对中文手机评论语料和中文酒店评论语料按照4:1的比例切分成训练集和测试集,实验数据如表1。

表1 实验数据

Tab. 1 Experimental data

语料	总计	正向	负向	训练集	测试集
酒店评论	1 500	750	750	1 200	300
手机评论	1 925	962	963	1 540	385

采用中科院的中文分词系统 NLPIR 对中文手机评论语料和中文酒店评论语料进行分词并词性标注。

2.2 评价方法

为了和方法 SCCR^[13]进行对比,本文采用正确率 α 评价中文评论情感分类的结果。正确率是一种常用的度量标准,同时考虑了正类和负类分类正确的概率,能体现两个类别的总体分类效果。

$$\alpha = \frac{N_{TP} + N_{TN}}{N_{TP} + N_{TN} + N_{FP} + N_{FN}}. \quad (7)$$

式中: N_{TP} 为将正类评论文本预测为正类的数目; N_{FN} 为将正类评论文本预测为负类的数目; N_{FP} 为将负类评论文本预测为正类的数目; N_{TN} 为将负类评论文本预测为负类的数目。

2.3 训练集规模实验

为了测试不同大小的训练集对情感分类准确性的影响,分别选择1/4的训练集、1/2的训练集、3/4的训练集及所有的训练语料进行实验。如图3和图4所示。选择N、V、Adj、Adv的结合NVAA作为特征,选择DF进行特征降维,抽取不同数量(150、200、250)的特征,采用布尔加权,使用SVM进行情感分类。

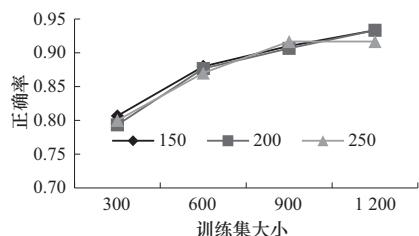


图3 训练集规模对分类正确率的影响(手机)

Fig. 3 Effect of the size of training set on classification accuracy (cell phone)

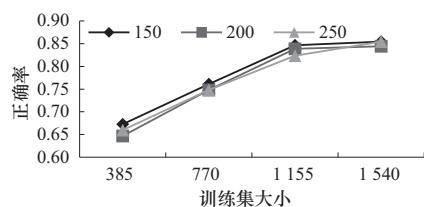


图4 训练集规模对分类正确率的影响(酒店)

Fig. 4 Effect of the size of training set on classification accuracy (hotel)

由实验结果可知,训练集规模对情感分类的结果会产生影响,在训练集规模较小时,随着训练集规模越大,分类的正确率越高,但是当训练集大小到达一定值后,再增大训练集规模,正确率提升很小或不再提升。原因是当训练集不充分时,获得的情感特征较少,因此可以适当扩充训练集的大小来提高准确率,以下实验都采用所有的训练语料。

2.4 特征选择效果实验

为了评估词嵌入相似性度量对扩展语义相似情感特征效果的影响,分别选择N、V、Adj、Adv、NVAA作为特征,以及NVAA+W2V,即基于词嵌入扩充情感特征,再用DF进行特征降维,SVM进行分类,特征维度分别设为150、200、250、500进行实验。如图5和图6所示。

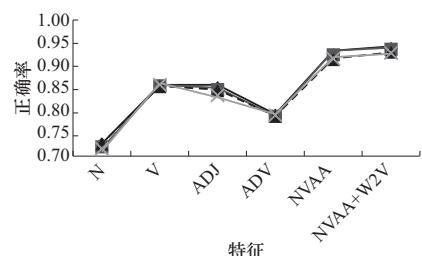


图5 特征选择实验(手机)

Fig. 5 Feature selection experiment (cell phone)

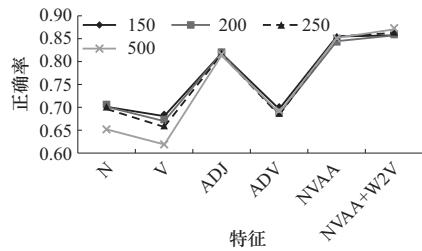


图6 特征选择实验(酒店)

Fig. 6 Feature selection experiment (hotel)

由实验结果可以看出NVAA+W2V的分类正确率最高,其次是NVAA和形容词,整体来说名词对情感分类的贡献比较少。原因是语料中,形容词“舒

适的”、“便利”等词的情感色彩强烈,而中性名词“功能”、“手机”、“酒店”等和副词“还”、“就”,以及动词“住”、“按”等很少包含感情色彩,形容词比名词、动词、副词更能反映评论者的情感。

基于词嵌入增加情感特征整体上提升了分类正确率。在酒店语料中,NVAA+W2V 作为特征正确率能达到 87.01%,比选择 NVAA 作为特征的分类正确率提升了 1.82%,在手机评论语料中,NVAA+W2V 作为特征正确率能达到 94.33%,提升了 1%。原因是从语料中提取的 NVAA 特征稀疏,而通过词嵌入计算相似度添加进新特征扩充了情感语义特征。由于词嵌入通过词语之间的语义信息计算词语相似度,添加进的新特征和原有特征语义上相近,增添了情感色彩,更能表达评论者的情感。基于词嵌入添加新情感特征优化了情感分类的效果。

2.5 特征降维效果实验

为了评估引入词语间的统计特征进行特征降维对情感分类性能的影响,选择 NAVV 作为中文评论语料的特征,选择 DF、IG、MI 及改进的方法 DF-TF、IG-TF、MI-DF 进行特征降维实验。

分别采用 DF、DF-TF、IG、IG-TF、MI、MI-DF 方法,计算 NVAA 特征中每个特征对应的特征降维方法权值,按照权值的大小从大到小排序,依次取排列前面的一定维度(150, 200, 250, 300, 500, 1 000)的特征。使用布尔权重,采用 SVM 分类器进行训练和测试,如图 7 和图 8 所示。

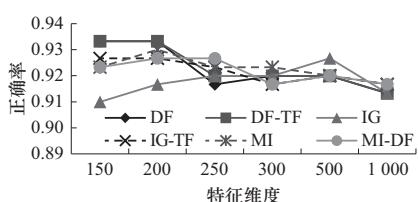


图 7 特征降维效果实验(手机)

Fig. 7 Experiment of effect on feature dimension reduction (cell phone)

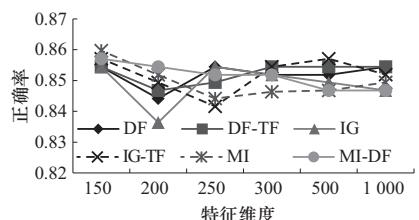


图 8 特征降维效果实验(酒店)

Fig. 8 Experiment of effect on feature dimension reduction (hotel)

由实验结果可知选择 DF、IG、MI 及改进的方法 DF-TF、IG-TF、MI-DF 分别进行特征降维对情感分类正确率的影响不是很大。使用 DF-TF 和 DF 进行特征降维获得的正确率相对较高,改进后的 DF-TF 整体效果最好。这是因为 DF 方法效果好,DF-TF 又考虑了词频,若一个特征出现在多个文档中,且在一个文档中出现多次,那它将具有较大的区分度。式(5)中经验参数 α 取 0.001,IG-TF 方法相对 IG 方法有近 1 个百分点的提升,这是因为 IG 算法只在乎特征项对分类结果的贡献度,不关注词频的影响,而 IG-TF 方法考虑了特征出现的频率,能取得更好的分类效果。MI-DF 方法相对 MI 方法进行情感分类的正确率提升较小,说明 DF 对 MI 的影响较小。MI 算法本身趋向于低频词,其分类效果比 DF 稍差,而特征数量增多后,分类效果趋于好转,主要是特征维度的增多包含了更多的高频特征。总的来说,DF-TF 方法由于其有效性和易于实现的优点,适合在大规模文本集中使用,处理了特征维度高的问题,更实用更方便。

2.6 情感分类效果实验

为了评估扩展语义相似情感词、引入词语间的统计特征和采用 Adaboost 分类模型的方法的整体效果,本实验在酒店、手机评论两个语料上与 SCCR 方法^[14-17]进行对比。SCCR 方法对中文在线评论进行情感分类,分类效果较好,可比性较强。实验选择 NVAA 作为初始特征,使用中文维基百科语料训练 Word2vec 生成词嵌入,词嵌入维度为 200 维。对于每一条评论,应用词嵌入计算其他词语与 NVAA 中每个词语的语义相似度,将相似度大于一定阈值的词语添加进 NVAA 特征集扩充情感语义特征,阈值的取值范围为 0~1,以 0.05 为间隔循环设置阈值,最终选取使得准确率最高的阈值。把扩充后的特征集作为最终特征集,利用引入词语间统计信息的特征降维方法 DF-TF 进行特征降维,采用布尔权重,将评论集表示成特征-评论的向量空间模型,最后采用由弱分类器加权集成的 Adaboost 分类模型进行情感分类。表 2 为情感分类对比实验。

实验结果表明,本方法整体上优于 SCCR,其中,酒店评论语料中,准确率达到 90.91%,比 SCCR 最好结果提升了 5.46%。手机评论语料中,准确率达到 93.67%,比 SCCR 最高结果提升了 0.34%,在特征数量 1 000 维时提升了 2%。原因是首先,本方法的泛化性比 SCCR 方法好,SCCR 方法

的泛化性较差,该方法在手机语料中的准确率较高,而在酒店评论语料中的准确率较低;其次,手机语料包含的属性较固定,所以手机语料中的特征重复较多,扩展语义相似的情感词并不能较多的拓展其情感特征,所以对手机评论语料的分类准确率提升较小。选择NVAA作为情感特征时情感语义信息不足,而结合词嵌入的特征项选择方法考虑了词语之间的语义信息,扩展语义相近的情感词更能反映评论者的情感。随着特征数量的增多,本方法在酒店语料中的分类准确率也逐渐增大,而在手机语料中的准确率先增大后平稳。原因是当特征数量相等时,酒店语料中包含的情感特征比手机语料少,特征数量越大使得酒店语料包含的情感信息越多。基于词嵌入相似性度量扩充情感特征,引入词语间统计特征进行特征降维和通过多个弱分类器加权构造Adaboost分类模型实现情感分类整体上提升了分类的性能。

表2 情感分类对比实验

Tab. 2 Comparative experiments of sentiment classification

特征数量	准确率/%			
	酒店		手机	
	SCCR	本方法	SCCR	本方法
150	85.45	87.01	93.33	92.33
200	84.42	87.01	93.33	93.67
250	85.45	87.53	91.67	93.00
300	85.19	87.28	92.00	92.33
500	85.19	89.61	92.00	93.33
1 000	85.45	90.91	91.33	93.33

3 结 论

针对文本情感分类时存在的情感语义特征利不足、特征降维效果欠佳等问题,提出了一种扩展语义相似情感词的文本情感分类方法。该方法首先用中文维基百科语料训练Word2vec中的Skip-gram模型生成词嵌入,基于词嵌入相似性度量扩展语义相似的词语为情感特征,再利用词语间的统计特征进行特征降维,最后通过多个弱分类器加权集成Adaboost分类模型进行情感正负倾向分类。实验结果表明,在酒店评论语料中,准确率能达到90.91%,相比对比方法提升了5.46%,在手机评论语料中,准确率能达到93.67%,提升了0.34%。该方法扩展语义相似情感词有利于丰富文本情感语义特征,引入词语间的统计特征有更好的特征降维效果,进一步提升了情感分类的正确率,有较好的实用性。

将来研究可以在以下3方面进行:①挖掘评论和服务、产品更深的情感语义信息;②衡量评论的情感强度;③在更多的领域上证明本方法的适用性。

参 考 文 献:

- [1] 卢兴. 基于统计方法的中文短文本情感分析[D]. 北京: 北京理工大学, 2016.
Lu Xing. Sentiment analysis of Chinese short text based on statistical method [D]. Beijing: Beijing Institute of Technology, 2016. (in Chinese)
- [2] Serrano-Guerrero J, Olivas J A, Romero F P, et al. Sentiment analysis:a review and comparative analysis of web services[J]. Information Sciences, 2015,311(5):18-38.
- [3] Taboada M, Brooke J, Tofiloski M, et al. Lexicon-based methods for sentiment analysis[J]. Computational Linguistics, 2011,37(2):267-307.
- [4] Duwairi R M, Marji R, Sha'ban N, et al. Sentiment analysis in arabic tweets[C]// 2014 5th International Conference on Information and Communication Systems(icics). [S. l.]: IEEE, 2014:1-6.
- [5] Yang A M, Zhou Y M, Lin J H. A method of Chinese texts sentiment classification based on Bayesian algorithm[C]// Applied Mechanics and Materials. [S. l.]: Trans Tech Publications, 2013,263:2185-2190.
- [6] Rao Y, Xie H, Li J, et al. Social emotion classification of short text via topic-level maximum entropy model[J]. Information & Management, 2016,53(8):978-986.
- [7] Cao Y, Chen Z, Xu R, et al. A joint model for Chinese microblog sentiment analysis[J]. ACL-IJCNLP, 2015, 2015:61-67.
- [8] Ying C, M Qi guang, Liu Jiachen, et al. Advance and prospects of AdaBoost algorithm[J]. Acta Automatica Sinica, 2013,39(6):745-758.
- [9] Bengio Y, Vincent P, Janvin C. A neural probabilistic language model[J]. Journal of Machine Learning Research, 2003,3(6):1137-1155.
- [10] Basu T, Murthy C A. Effective text classification by a supervised feature selection approach[C]// International Conference on Data Mining Workshops. [S. l.]: IEEE Computer Society, 2012:918-925.
- [11] Xu Y, Wang B, Li J T, et al. An extended document frequency metric for feature selection in text categorization [C]// Asia Information Retrieval Conference on Information Retrieval Technology. [S. l.]: Springer-Verlag, 2008:71-82.

(下转第 1176 页)

- model for radiated immunity measurements [J]. IEEE Transactions on Electromagnetic Compatibility, 2013, 55(4):683–691.
- [2] IEC. IEC 61000 – 4 – 21: Electromagnetic compatibility (EMC) – Part4 – 21: testing and measurement techniques–Reverberation chamber test methods [S]. [S. l.]: IEC, 2011.
- [3] Arnau L R T. Elliptic stochastic fields in reverberation chambers [J]. IEEE Transactions on Electromagnetic Compatibility, 2016, 58(1):11–21.
- [4] Lee Y S, Kwon J H, Park S K. Extension of a plane wave integral representation for fields in lossy reverberation chambers [J]. Electromagnetics, 2015, 35:249–259.
- [5] Gros J B, Legrand O, Mortessagne F, et al. Universal behaviour of a wave chaos based electromagnetic reverberation chamber [J]. Wave Motion, 2014, 51:664–672.
- [6] Selemanni K, Gros J B, Richalot E, et al. Comparison of reverberation chamber shapes inspired from chaotic cavities [J]. IEEE Transactions on Electromagnetic Compatibility, 2015, 57(1):3–11.
- [7] Tsigros C, Piette M, Vandenbosch G, et al. Radiated immunity in reverberation and semianechoic rooms: conditions for equivalence [J]. IEEE Transactions on Electromagnetic Compatibility, 2013, 55(2):222–230.
- [8] 贾锐, 王庆国, 王树桥, 等. 混响室条件下辐射敏感度表征方法研究 [J]. 北京理工大学学报, 2016, 36(1):100–104.
- Jia Rui, Wang Qingguo, Wang Shuqiao, et al. Research on the characterization of susceptibility threshold in reverberation chamber [J]. Transactions of Beijing Institute of Technology, 2016, 36(1):100–104. (in Chinese)
- [9] Hu Dezhou, Wei Guanghui, Pan Xiaodong, et al. Investigation of the radiation immunity testing method in reverberation chambers [J]. IEEE Transactions on Electromagnetic Compatibility, 2017, 59(6):1791–1797.
- [10] Wilson P F, Hill D A, Holloway C L. On determining the maximum emissions from electrically large sources [J]. IEEE Transactions on Electromagnetic Compatibility, 2002, 41(1):79–86.
- [11] Amador E, Miry C, Bouye N. Compatible susceptibility measurements in fully anechoic room and reverberation chamber [C]// Proc. of the 2014 International Symposium on Electromagnetic Compatibility (EMC Europe 2014). Sweden: [s. n.], 2014:860–865.

(责任编辑:刘芳)

(上接第 1162 页)

- [12] Wang H, Yin P, Zheng L, et al. Sentiment classification of online reviews: using sentence-based language model [J]. Journal of Experimental & Theoretical Artificial Intelligence, 2014, 26(1):13–31.
- [13] Wang H, Zheng L. Sentiment classification of Chinese online reviews: a comparison of factors influencing performances [J]. Enterprise Information Systems, 2016, 10(2):228–244.
- [14] Li W, Li Y, Wang Y. Chinese microblog sentiment analysis Based on sentiment features [C]// Asia-Pacific Web Conference. Cham: Springer, 2016:385–388.

- [15] Gustafsson P, Granholm J. The quest for the abnormal return: a study of trading strategies based on twitter sentiment [D]. Umeå, Sweden: Umeå Universitet, 2017.
- [16] Wang H, Guo K. The impact of online reviews on exhibitor behaviour: evidence from movie industry [J]. Enterprise Information Systems, 2016, 11(10):1–17.
- [17] Wang H, Song Y, Du Z, et al. Evaluation of service quality for express industry through sentiment analysis of online reviews [J]. Beijing Gongye Daxue Xuebao, 2017, 43(3):402–412.

(责任编辑:刘芳)