基于双语词典的微博多类情感分析方法

栗雨晴[1,2],礼 欣[1,2],韩 煦[1],宋丹丹[1,2],廖乐健[1,2]
(1. 北京理工大学计算机学院,北京 100081;2. 北京市海量语言信息处理与云计算应用工程技术研究中心,北京 100081)

摘 要: 现有微博文本情感分析方法多面向单一语种语料,如:中文语料.但是,中英文搭配使用的表达习惯已逐渐成为个体意见表达的重要形式.本文提出一种基于双语词典的多类情感分析方法,通过构建双语多类情感词典对微博文本进行多分类语义倾向性分析,以便更准确有效捕捉群体意见,及时发现社会舆论倾向.通过与多数投票算法、支持向量机算法、基于余弦距离的K近邻分类算法相比,本文提出的基于双语词典的多类情感分析模型具有良好的分类效果,其在分类准确率、F1值等方面都有明显提高.

关键词: 双语语义倾向性分析;半监督高斯混合模型;相对熵;情感词典
中图分类号: TP391;H085.5 **文献标识码:** A **文章编号:** 0372-2112 (2016)09-2068-06
电子学报URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2016.09.007

A Bilingual Lexicon-Based Multi-class Semantic Orientation Analysis for Microblogs

LI Yu-qing[1,2], LI Xin[1,2], HAN Xu[1], SONG Dan-dan[1,2], LIAO Le-jian[1,2]
(1. *School of Computer Science, Beijing Institute of Technology, Beijing* 100081, *China*;
2. *Beijing Engineering Application Research Center of High Volume Language Information Processing and Cloud Computing, Beijing* 100081, *China*)

Abstract: Most of the existing Weibo sentiment analysis focuses on monolingual corpus like Chinese. However, a mixed use of Chinese and English becomes a popular form of expression. To better capture the social attention on public events, this paper proposes a bilingual lexicon based multi-class semantic orientation analysis for bilingual microblogs. We compare our proposed methodologies with majority vote, support vector machine (SVM) and K-nearest neighbor (KNN) by using cosine similarity which are competitive baseline methods. The experimental results show that our proposed methods outperform the three approaches we mentioned in terms of the accuracy and F1 score.

Key words: bilingual semantic orientation analysis; semi-supervised gaussian mixture model (Semi-GMM); Kullback-Leibler divergence; sentiment lexicon

1 引言

随着社交媒体平台的兴起和广泛使用,针对社交网络数据的自然语言处理已成为当前研究热点并囊括多种前沿课题.

目前,一些情感分析方面的工作主要针对单一语种文本情感倾向进行统计分析,但中英文搭配使用或纯英文书写已逐渐成为个体情感表达的重要形式.在本文中我们通过利用大量语料、已有知识库、词汇相似性计算模型构建英汉双语情感词词典,进而对微博文本进行向量化处理.本文利用半监督高斯混合模型分类算法(Semi-GMM, Semi-supervised Gaussian Mixture Model)和基于对称相对熵的K近邻算法(KNN-KL, K-Nearest Neighbor-symmetric Kullback-Leibler divergence)对微博文本进行情感分类.实验证实,半监督高斯混合模型分类算法鲁棒性强,并且分类准确率不受训练集文本规模大小的影响,而基于对称相对熵的K近邻算法(KNN-KL)在训练数据充分的情况下,可以取得更高的分类准确率.

2 相关工作

目前国内外对于文本情感倾向性判定主要有基于语料库和基于词典两种方法.总体来看,使用情感词典及与其相关联信息对文本进行情感判别效果更加精

收稿日期:2015-02-03;修回日期:2015-07-20;责任编辑:覃怀银
基金项目:国家重点基础研究发展规划(973计划)项目(No.2013CB329605);国家自然科学基金(No.61300178)

准[1]. 针对微博上大量中英双语混合文本的出现,我们通过构建双语情感词典以提高情感倾向分析的准确性.

在文献[2,3]中,作者提出跨语言混合模型,利用平行语料库提高词典覆盖率,通过最大化生成语料库的似然值对未标注词语进行情感极性标注,进而扩展词典. 但是,利用平行语料库的方式进行文本情感分类对平行语料库质量、规模要求很高. 微博文本内容简短、词汇复杂多变不利于平行语料库的构建. 因此,本文首先对大规模语料进行统计分析,预先对具有代表性的词汇进行人工标注选为种子词汇,再利用已有情感词汇知识库、语义相似度计算模型或层次结构模型等方法对双语情感词典进行扩充. 在构建词典的过程中我们利用新浪微博消息文本、中英文种子词集结合双语料相似度计算模型构建情感词典.

在文献[4]中作者指出采用机器学习方法比简单统计褒义和贬义情感词汇个数具有更好的分类效果,并提出将情感词典同监督学习算法相结合以实现更高的文本分类精度. 在文献[5]中,作者提出一种反应公众对社会事件关注的五分类模型(社会关爱、高兴、悲伤、愤怒、恐惧). 本文结合上述文本情感类别,提出基于半监督高斯混合模型等一系列动态学习算法对中文及中英双语微博文本进行情感倾向性分类.

3　文本情感分类系统

本章将从情感词典的构建、文本情感倾向性分类、文本的向量表示以及文本情感分类算法设计四个方面的研究工作进行介绍. 文本情感分类系统的整体框架如图 1 所示:

3.1　情感词典

中英文搭配使用已成为个体表达的流行趋为进一步说明加入英文情感词典的必要性,我们在图 2 中展示微博用户发布的两则博文,图中可以看出,具有双语表述习惯的用户在谈及某一话题时,惯用英语情感词汇进行情感表达.

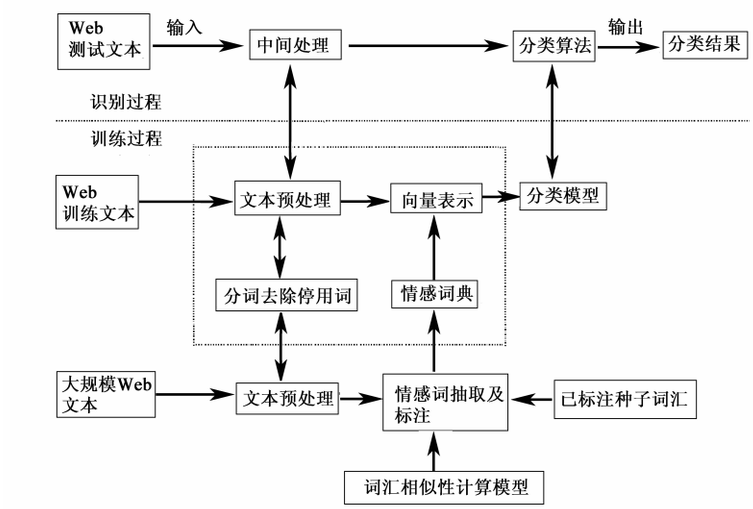

图1　微博文本感情分类系统结构框架

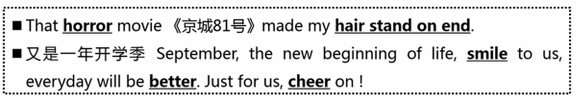

图2　双语微博文本

为建立双语情感词典,首先我们从新浪微博中收集大量具有情感倾向的语料,并从语料集中提取出具有情感倾向的高频词汇. 之后,应用已有知识库(HowNet[6]、WordNet[7]、NTUSD[8])对情感词典进行扩展. 在已有知识库中(HowNet[6]、WordNet[7])每个词汇 $v_b(b \in Z^+)$ 可以通过多个概念 $S_{ba}(a \in Z^+)$ 进行描述,每个概念又是以义原为基础通过知识库表述语言进行定义,且每个概念 S_{ba} 含有多个义原 $p_{at}(t \in Z^+)$ 对其进行解释. 对于中文词汇间的语义相似性,本文采用 HowNet 词汇相似度计算方法[9],其定义如式(1)、式(2)所示:

$$\mathrm{sim}(v_1,v_2) = \max_{a_1=1\cdots n,a_2=1\cdots m} \mathrm{sim}(S_{1a_1},S_{2a_2}) \tag{1}$$

$$\mathrm{sim}(S_{1a_1},S_{2a_2}) = \frac{1}{t_1 t_2}\sum_{l_1=1}^{t_1}\sum_{l_2=1}^{t_2}\frac{1}{2^{t_1-l_1+1}}\mathrm{sim}(p_{a_1l_1},p_{a_2l_2}) \tag{2}$$

其中,t_1,t_2 分别表示 S_{1a_1},S_{2a_2} 两个概念含有的义原数目,并选取两个词之间的最大概念描述相似度作为两个词的相似度.

而对于英文词汇间的语义相似性,我们利用 WordNet 中的 Lesk 方法对词汇之间的关联度进行度量. 在 Wordnet 中的每一个概念(word sense)都是通过一个短注释进行定义的. Lesk 方法通过寻找和计算两个概念的注释的交叉部分进而计算两词汇之间的相似度 sim(v_1,v_2). 本文采用 NLTK 中给出的 Lancaster 和 WordNet Lemmatizer 两种方式对英文词汇进行词形变化和词干提取. 除传统情感词外,我们还在情感词典中引入了网络语言和表情符号. 综上所述,本文所构建的中文情感词汇共计 7590 个,英文情感词汇共计 421 个,网络词汇 613 个,常用表情符号 101 个.

3.2　文本的向量表示

根据构建的情感词典我们从中选取部分词汇进行人工标注作为 5 类情感的种子词汇. 种子词集 $A_{\_\mathrm{seedset}} = \{P_C,P_J,P_B,P_A,P_F\}$,其中 P_C,P_J,P_B,P_A,P_F 分别代表各类情感(社会关爱、高兴、悲伤、愤怒、恐惧)的子集. 其中"社会关爱"类别的引入旨在更准确有效的捕捉、辨别群体意见[5],而对于不在种子集合中的情感词,我们则利用式(3)中所给定义将其分类.

$$\Psi(v)=\arg\max_{\{P_C,P_J,P_B,P_A,P_F\}}\left(\frac{1}{K_1}\sum_{k_1=1}^{K_1}\mathrm{sim}(v,P_{C\_k_1}),\frac{1}{K_2}\sum_{k_2=1}^{K_2}\mathrm{sim}(v,P_{J\_k_2}),\frac{1}{K_3}\sum_{k_3=1}^{K_3}\mathrm{sim}(v,P_{B\_k_3}),\frac{1}{K_4}\sum_{k_4=1}^{K_4}\mathrm{sim}(v,P_{A\_k_4}),\frac{1}{K_5}\sum_{k_5=1}^{K_5}\mathrm{sim}(v,P_{F\_k_5})\right)\tag{3}$$

其中 K_1,K_2,K_3,K_4,K_5 为各类情感子集中种子词汇的数目. $\Psi(v)$ 表示非种子词汇所属情感类别,取决于与各类情感子集平均相似度的最大值. 而对于微博消息中常出现的网络词汇则采用多人人工标注的方式对其进行分类. 最终建立的中英双语五类情感词典涵盖“社会关爱”类词汇 971 个、“高兴”类词汇 2731 个、“悲伤”类词汇 2289 个、“愤怒”类词汇 1458 个、“恐惧”类词汇 1276 个.

本文采用 ICTCLAS 分词系统 (http://ictclas.nlpir.org/)对中文文本进行词汇识别,而对于英文文本则根据空格进行词汇识别. 对一条微博消息文本进行分此后,对其进行去停用词处理,如:“的”、“a”、“the”等.

对微博消息文本进行上述处理之后便可依照多分类情感词典对其进行文本向量化表示. 设 $D=\{\boldsymbol{d}_1,\boldsymbol{d}_2,\cdots,\boldsymbol{d}_n\}$ 是所有微博消息文本的集合,其中 $\boldsymbol{d}_i$ 是本文集合中第 i 条文本的向量表示. 则对于任一条微博文本 $\boldsymbol{d}_i=[\omega_{iC},\omega_{iJ},\omega_{iB},\omega_{iA},\omega_{iF}]^{\mathrm{T}}$ 其中 $\omega_{iC},\omega_{iJ},\omega_{iB},\omega_{iA},\omega_{iF}$ 表示微博消息文本中包含各类情感词的个数,因此每条微博消息均以 5 维向量表示.

3.3 算法设计

本节将详细介绍本文提出的两种文本情感多分类模型——半监督高斯混合模型分类算法(Semi-GMM)和基于对称相对熵的 K 近邻算法(KNN-KL).

3.3.1 半监督高斯混合模型(Semi-GMM)情感分类算法

高斯混合模型学习,即是对各个高斯模型加概率密度的估计和权重(π_k)进行最大似然估计的过程. 本文采用半监督高斯混合模型对文本进行分类,首先通过已标记微博消息文本学习高斯混合模型,然后以该模型参数和已标记样本的概率分布作为高斯混合模型的参数初值对已有模型进行迭代学习.

半监督高斯混合模型是一个自训练算法,在每一次迭代训练的过程中,已标注样本集合(L)通过不断在未标注样本集合(U)中选择表现良好的样本加入,更新标注样本集. 根据新的标注集合不断对混合高斯模型进行学习,直至算法收敛或未标注集合为空. 半监督高斯混合模型情感分类算法伪代码如算法 1 所示:

算法 1 半监督高斯混合模型情感分类算法

输入:小规模已标注微博文本集合,高斯混合模型
输出:$\Theta^{(q)}$

```
q←0
Θ^(0)←arg max_θ P(θ|L)
while U! = NULL or ‖Q(θ^(q+1),θ^(q)) − Q(θ^(q),θ^(q))‖ > ε
E_step:
do  γ_jk ← α_k φ(u_j | θ_k) / Σ_{k=1}^{K} α_k φ(u_j | θ_k)
    (j,k) ← arg max_(j,k) {γ_jk | u_j}
    L ← L ∪ u_j
    U ← U − u_j
M_step:
    Θ^(q) ← arg max_θ P(θ | Γ^(q), L)
    q ← q + 1
```

其中 L 是已标注微博文本集合,U 是未标注微博文本集合,K 表示类别个数. $\varphi(u|\theta)$ 是高斯概率密度函数,$\theta=(\mu_k,\sigma_k^2)$.

3.3.2 基于对称相对熵的 K 近邻情感分类算法

K 近邻分类算法(KNN, K-Nearest Neighbor)[10] 是指一个样本所属类别取决于特征空间中最邻近的样本中大多数所属类别. 在本文中我们采用相对熵对文本情感相似性进行度量. 相对熵是对相同事件空间里的两个概率分布(P 和 Q 的)的非对称性度量,记为 $D_{\mathrm{KL}}(P\|Q)$. 因此对 3.3 节中提出的文本向量表示进行归一化,如式(4)所示,归一化后的文本向量记为 $\boldsymbol{T}_i$,其中 W 为文本包含各类情感词的个数总和.

$$\boldsymbol{T}_i=\langle\omega_{iC}/W,\omega_{iJ}/W,\omega_{iB}/W,\omega_{iA}/W,\omega_{iF}/W\rangle\tag{4}$$

微博消息文本 $\boldsymbol{T}_i$ 与 $\boldsymbol{T}_j$ 之间的距离定义如式(5)所示:

$$D_{\mathrm{KL}}(\boldsymbol{T}_i\|\boldsymbol{T}_j)=\sum_k t_{ik}\log_2\frac{t_{ik}}{t_{jk}}\tag{5}$$

由于传统相对熵具有非对称性,因此在度量概率分布 P 和 Q 的差别时,P 表示数据的真实分布,Q 表示 P 的近似分布. 因此,在计算文本之间的距离时,$\boldsymbol{T}_i$ 为已标记文本的归一化向量表示,$\boldsymbol{T}_j$ 则为未标记文本的归一化向量表示. t_{ik}但是这种非对称性计算形式忽略了 P 对于 Q 的近似分布. 为了改进传统相对熵计算的非对称性,本文采用的相对熵计算公式[11] 定义如式(6)所示:

$$\frac{1}{D_{\mathrm{KL}}(\boldsymbol{T}_i\|\boldsymbol{T}_j)}=\frac{1}{\sum_k t_{ik}\log_2\frac{t_{ik}}{t_{jk}}}+\frac{1}{\sum_k t_{jk}\log_2\frac{t_{jk}}{t_{ik}}}\tag{6}$$

4 实验结果及相关分析

4.1 多种文本情感分类算法比较

本实验根据 3.1 节中构建的中文情感词典,选取多种机器学习分类算法进行比较. 使用新浪微博提供的 API 抓取 7170 条中文微博文本信息作为实验数据. 并

邀请25位研究自然语言方向的学生依照5类情感对文本进行人工类别标注,进而使得文本的情感类别取决于多数人选取的情感类别. 语料在各情感类别中的分布情况如表1所示:

表1　微博文本在5类情感类别中的分布

	社会关爱	高兴	悲伤	愤怒	恐惧
文本数量	1300	2100	1340	1310	1120
所占比例	18.1%	29.3%	18.7%	18.3%	15.6%

针对上述微博文本我们采用多种分类模型对文本进行情感分类,实验详细设计与结果分析如下所述.

我们从中选取3170条微博作为测试集,其中表达社会关爱的微博文本500条,表达高兴的微博文本1300条,表达悲伤的微博文本540条,表达愤怒的微博文本510条,表达恐惧的微博文本320条. 训练集则从余下4000条中选取1000至4000条微博不等.

(1)我们首先对基于非对称相对熵的K近邻分类算法,如式(5)所示和基于对称相对熵的K近邻分类算法,如式(6)所示进行比较,实验结果如表2所示.

结果表明,尽管基于对称相对熵的K近邻分类算法依照本文所示训练文本优势并不明显,但考虑到基于对称相对熵的K近邻分类算法可消除不同训练集导致的算法准确率差异,进而提高分类算法的高鲁棒性. 因此,在之后的多种机器学习分类算法的比较中,我们仅选用基于对称相对熵的K近邻分类算法参与比较.

表2　基于不同距离度量算法的K近邻分类算法在不同训练集规模下的准确率比较

训练集文本数量	非对称相对熵距离度量算法	对称相对熵距离度量算法
1000	76.1%	76.2%
1500	79.3%	79.3%
2000	80.8%	80.9%
2500	81.5%	81.8%
3000	82.7%	83.2%
3500	84.0%	84.3%
4000	84.7%	85.1%

(2)多模型分类结果的比较

我们选用多数投票算法(Majority Vote)、支持向量机算法(SVM)、基于余弦距离的K近邻分类算法(KNN-Cosine)同本文中提出的半监督高斯混合模型分类算法(Semi-GMM)和基于对称相对熵的K近邻算法(KNN-KL)进行比较. 比较结果如图3所示:

从图3可以看出当训练集文本规模为4000条时,KNN-KL准确率最高达到85.1%. 当选用相同最近邻数时,采用对称相对熵进行文本距离度量比采用余弦距离进行文本距离度量分类效果更好. 但随着训练集文本数目下降到1000条,采用KNN-KL的准确率下降了8.9%,而Semi-GMM仅下降了2.9%. 这也进一步证实了Semi-GMM更加适合在训练集规模较小时使用,而KNN这种全监督学习算法容易被选取邻居数目左右,影响分类效果.

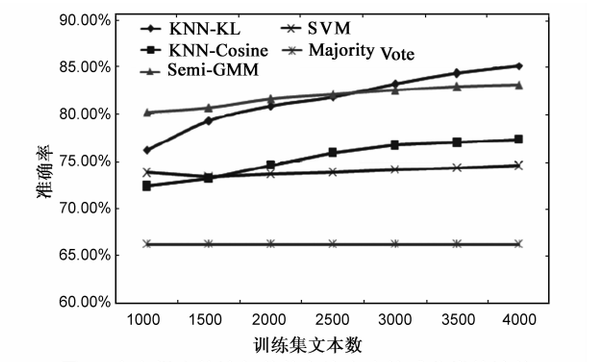

图3　本文提出算法与其他主流文本情感分类算法的准确率比较

表3　在不同训练集规模下,基于Semi-GMM和KMM-KL的文本分类准确率

分类算法	训练集文本数量	社会关爱	高兴	悲伤	愤怒	恐惧
KNN-KL	4000	90.6%	82.1%	80.7%	83.3%	98.8%
	1000	74.8%	76.5%	69.6%	76.3%	88.1%
Semi-GMM	4000	62.3%	81.1%	77.4%	81.8%	99.1%
	1000	85.8%	78.1%	74.4%	80.6%	91.6%

表4　在不同训练集规模下,基于Semi-GMM和KMM-KL的文本分类F1值

训练集文本数量	分类算法	社会关爱	高兴	悲伤	愤怒	恐惧
4000	Semi-GMM	69.1%	87.2%	78.0%	89.7%	93.1%
	KNN-KL	73.0%	88.6%	81.0%	90.3%	96.6%
1000	Semi-GMM	65.9%	85.1%	73.8%	89.0%	88.9%
	KNN-KL	63.8%	82.1%	68.9%	85.3%	89.8%

在不同文本训练集规模下,Semi-GMM和KNN-KL的F1值如表4所示,这也进一步证实了Semi-GMM在小规模训练集下的分类优势.

4.2　双语微博文本情感分类实验

类似的,我们使用新浪提供的API抓取7000条双语微博文本信息. 并邀请25位研究自然语言方向的学生依照5类情感对文本进行人工类别标注,情感类别语料在各情感类别中的分布情况如表5所示:

表5　微博文本在5类情感类别中的分布

	社会关爱	高兴	悲伤	愤怒	恐惧
文本数量	1200	1750	1460	1300	1290
所占比例	17.1%	25.0%	20.9%	18.6%	18.4%

针对上述双语微博文本我们采用多种分类模型对文本进行情感分类,实验详细设计与结果分析如下

所述.

(1)多模型分类结果比较

我们从中选取3000条微博作为测试集,其中表达社会关爱的微博文本400条,表达高兴的微博文本950条,表达悲伤的微博文本660条,表达愤怒的微博文本500条,表达恐惧的微博文本490条.训练集则从余下4000条中选取1000至4000条微博不等.

我们选用仅使用中文情感词典作感词识别的半监督高斯混合模型分类算法(Semi-GMM(Ch.))和基于对称相对熵的K近邻算法(KNN-KL(Ch.))同使用中英文情感词典相结合进行情感词识别的多数投票算法(Majority Vote(Ch.+Eng.))、SVM(Ch.+Eng.)算法、基于余弦距离的K近邻分类算法(KNN-Cosine(Ch.+Eng.))以及本文提出的半监督高斯混合模型分类算法(Semi-GMM(Ch.+Eng.))和基于对称相对熵的K近邻算法(KNN-KL(Ch.+Eng.))进行比较.比较结果如图4所示:

如图4所示,利用中英文情感词典相结合进行情感词识别的文本情感分类算法准确率明显高于单一利用中文情感词典进行情感词识别的文本情感分类算法,进一步证实了我们建立的双语情感词词典的有效性.当训练集微博文本下降到1000条时,Semi-GMM(Ch.+Eng.)的分类准确率最高达到了68.3%.

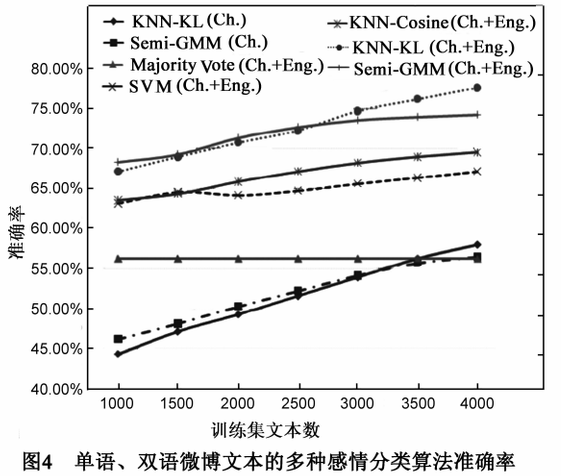

图4 单语、双语微博文本的多种感情分类算法准确率

表6 不同训练集规模下,基于Semi-GMM和KMM-KL的文本分类准确率

分类算法	训练集文本数量	社会关爱	高兴	悲伤	愤怒	恐惧
KNN-KL	4000	66.7%	84.7%	68.2%	81.2%	81.1%
	1000	53.2%	77.8%	58.2%	71.4%	67.5%
Semi-GMM	4000	62.3%	82.9%	65.2%	78.8%	76.5%
	1000	54.6%	78.8%	59.4%	72.9%	68.8%

表6和表7给出了当文本训练集规模不同时,Semi-GMM和KNN-KL针对文本进行5类情感识别的准确率.在文本训练集规模下降到1000时,Semi-GMM的F1值大于KNN-KL的F1值,这也进一步证实了文本中出现不同语种的文字不会对Semi-GMM的稳定性造成影响,并且在小规模训练集下Semi-GMM更具分类优势.

表7 不同训练集规模下,基于Semi-GMM和KMM-KL的文本分类F1值

训练集文本数量	分类算法	社会关爱	高兴	悲伤	愤怒	恐惧
4000	Semi-GMM	64.6%	84.7%	68.3%	75.7%	67.9%
	KNN-KL	69.2%	86.3%	71.3%	78.4%	72.1%
1000	Semi-GMM	56.9%	80.5%	63.0%	67.8%	61.1%
	KNN-KL	55.6%	79.7%	61.7%	66.4%	59.7%

(2)平行语料 vs. 双语情感词典

我们利用平行语料库方式对文本进行预处理——通过调用百度翻译API将双语微博文本信息全部翻译为中文单一语料文本.针对于上述构建完成的平行语料文本集,我们选用中文情感词典作情感词识别的半监督高斯混合模型分类算法(Semi-GMM(Ch.))和基于对称相对熵的K近邻算法(KNN-KL(Ch.))对文本就行情感分类.并与本文提出的基于双语情感词词典的半监督高斯混合模型分类算法(Semi-GMM(Ch.+Eng.))和基于对称相对熵的K近邻算法(KNN-KL(Ch.+Eng.))进行比较.

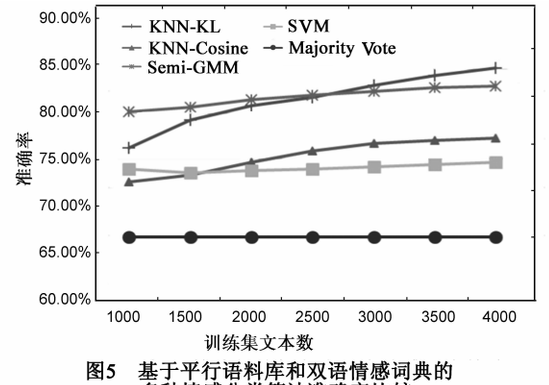

图5 基于平行语料库和双语情感词典的多种情感分类算法准确率比较

如图5所示,在选用相同分类模型的前提下,利用中英文情感词典相结合进行情感词识别的文本情感分类算法准确率明显高于利用平行语料库方式进行文本预处理的文本情感分类算法.由于情感词汇语义复杂,上述实验也印证了多类别情感词典的构建不适宜采用平行语料库的方式,证实了我们构建的双语情感词词典对于多类情感识别的有效性.

(3)微博文本英文字符所占比重对情感分类的影响

我们通过实验分析了微博文本中英文字符所占比重(新浪微博中每两个英文字符算为一字)对本文提出的情感分类算法的影响. 我们从 7000 条双语微博文本信息中随机选取其中 3000 条双语微博文本作为训练集. 测试集则从余下 4000 条中选取,其中英文字符所占比重小于 30% 的文本共计 1105 条(27.62%),英文字符所占比重介于 30% 至 70% 的文本共计 2170 条(54.25%),英文字符所占比重大于 70% 的文本共计 725 条(18.13%). 实验结果如表 8 所示:

表 8　不同英文字符占比测试集下的文本分类准确率比较

分类算法	英文字符占比	分类准确率
KNN-KL	<30%	74.7%
	30% ~70%	72.9%
	>70%	73.2%
Semi-GMM	<30%	73.6%
	30% ~70%	72.4%
	>70%	73.1%

结果表明,本文提出的情感分类算法的高准确率不受文本英文字符比重的影响. 这也进一步证明了我们建立的双语情感词词典的有效性以及分类模型的强鲁棒性.

5　结论

中英文搭配使用的表达习惯已成为社交网络个体、群体意见表达的重要形式. 本文使用新浪微博消息文本和已有知识库构建了双语情感词典. 为进一步加强面向语义分类器的性能,本文提出了半监督高斯混合模型和基于相对熵的 K 近邻算法对文本进行情感分类. 实验结果表明,本文提出的基于双语情感词典的情感分类方法的准确率和综合评价指标(F1 值)均高于传统的分类方法. 特别是半监督高斯混合模型分类算法在小规模训练集下的分类效果明显优于其他方法.

参考文献

[1] Melville P, Gryc W, Lawrence R D. Sentiment analysis of blogs by combining lexical knowledge with text classification[A]. Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining[C]. New York: ACM SIGKDD Explorations Newsletter, 2009. 1275 – 1284.

[2] Wan X. Bilingual co-training for sentiment classification of Chinese product reviews[J]. Computational Linguistics, 2011, 37(3): 587 – 616.

[3] Meng X, Wei F, Liu X, et al. Cross-lingual mixture model for sentiment classification[A]. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1[C]. Stroudsburg: Association for Computational Linguistics, 2012. 572 – 581.

[4] Pang B, Lee L. Opinion mining and sentiment analysis[J]. Foundations and Trends in Information Retrieval, 2008, 2(1-2): 1 – 135.

[5] Li Y, Li X, Li F, et al. A lexicon-based multi-class semantic orientation analysis for microblogs[A]. Web Technologies and Applications[C]. Cham: Springer International Publishing, 2014. 81 – 92.

[6] Dong Z, Dong Q. HowNet and the Computation of Meaning[M]. Singapore: World Scientific, 2006.

[7] Miller G A. WordNet: a lexical database for English[J]. Communications of the ACM, 1995, 38(11): 39 – 41.

[8] Hu M, Liu B. Opinion extraction and summarization on the web[A]. Proceedings of the 21st National Conference on Artificial Intelligence(AAAI 2006)[C]. California: AAAI Press, 2006. 1621 – 1624.

[9] Zhu Y L, Min J, Zhou Y, et al. Semantic orientation computing based on HowNet[J]. Journal of Chinese Information Processing, 2006, 20(1): 14 – 20.

[10] Chen J, Xue N, Palmer M S. Using a smoothing maximum entropy model for Chinese nominal entity tagging[A]. Natural Language Processing-IJCNLP 2004[C]. Heidelberg: Springer-Verlag Berlin Heidelberg, 2004. 493 – 499.

[11] Seghouane A K, Amari S I. The AIC criterion and symmetrizing the Kullback-Leibler divergence[J]. IEEE Transactions on Neural Networks, 2007, 18(1): 97 – 106.

作者简介

栗雨晴　女,1991 年 7 月出生于北京市. 现为北京理工大学硕士研究生. 主要研究方向为社交网络、文本情感分析.
E-mail: liyqyimy@163.com

礼　欣(通讯作者)　女,1980 年 4 月出生于黑龙江省佳木斯市. 2001 和 2004 年分别获得吉林大学计算机学院工学学士和硕士学位,2009 年获香港浸会大学计算机博士学位. 目前就职于北京理工大学计算机学院,主要从事数据挖掘、机器学习、无线传感网、车联网、社交网络分析和移动计算等方面的研究.
E-mail: xinli@bit.edu.com