文章编号：1003-0077(2017)02-0187-07

大规模情感词典的构建及其在情感分类中的应用

赵妍妍[1]，秦 兵[2]，石秋慧[2]，刘 挺[2]

(1. 哈尔滨工业大学 媒体技术与艺术系，黑龙江 哈尔滨 150001；
2. 哈尔滨工业大学 计算机学院，黑龙江 哈尔滨 150001)

摘 要：以微博为代表的社会媒体的飞速发展为情感分析方向带来巨大的资源，同时也对情感分析算法的性能提出了更大的挑战。其中，现有的情感词典尤其是中文情感词典规模不足是影响情感分析性能的一个重要因素。为此，该文基于海量的微博数据，使用简单的文本统计算法，构建了一个十万词语/词组的大规模情感词典。我们以情感分析的基础任务——情感分类为例，将大规模情感词典作为特征用于该任务上，实验结果表明大规模词典有助于情感分类性能的提高。

关键词：情感词典；情感分析；情感分类；微博

中图分类号：TP391　　**文献标识码**：A

Large-scale Sentiment Lexicon Collection and Its Application in Sentiment Classification

ZHAO Yanyan[1], QIN Bing[2], SHI Qiuhui[2], LIU Ting[2]

(1. Department of Media Technology and Art, Harbin Institute of Technology, Harbin, Heilongjiang 150001, China;
2. Department of Computer Science and Technology, Harbin Institute of Technology, Harbin, Heilongjiang 150001, China)

Abstract: Rapid development of social media, such as Micro-blog, brings lots of information as well as challenges for sentiment analysis. The limited size of Chinese sentiment lexicon is one critical influence on the performances of sentiment analysis. This paper proposes a simple statistical method to mine large amounts of sentiment words or phrases to construct a large scale 100,000 words/phrases from microblogs. We apply this large-scale lexicon to Chinese microblog sentiment classification, and the results confirm a clear performance improvement.

Key words: sentiment lexicon; sentiment analysis; sentiment classification; chinese microblog

1 引言

社会媒体，例如，论坛、博客、微博的出现，将以往媒体一对多的传播方式改变为用户参与多对多的“对话”。因此，随着社会媒体的深入发展和用户的不断参与，用户在互联网上产生的数据(User Generated Content，UGC)呈爆发式增长。根据新浪的统计数据，目前用户每日发博量约一亿条。UGC 信息多是评论信息，表达了网民对某个事件、某个人或某款产品的情绪和情感倾向性。UGC 信息量的剧增迫切需要情感分析技术帮助用户快速获取和整理这些相关评价信息以了解大众舆论。因此，海量数据给网民的生活提供丰富的信息，为情感分析方向带来了巨大的数据资源的同时，也对情感分析算法的性能提出了更大的挑战。

情感词典是情感分析领域重要的资源，几乎每个涉及到情感分析的研究任务和研究算法都会应用到情感词典[1-2]。在英文词典方面具有代表性的有 GI① (General Inquirer)英文情感词典，该词典收集了1 914 个褒义词和 2 293 个贬义词，并为每个词语按照极性、强度、词性等打上不同的标签，便于情感

① http://www.wjh.harvard.edu/~inquirer/

收稿日期：2016-09-27　**定稿日期**：2016-10-15

基金项目：中国博士后科学基金(2012M520740，2013T60373，2012M520142)

分析任务中的灵活应用。此外，还有 Opinion Lexicon①，包含约 6 800 个褒义词和贬义词。在中文情感词典方面，比较有代表性的有 HowNet 情感词典，包含 9 193 个褒义词和贬义词。然而，现有的情感词典在情感分析任务的使用中存在三点不足。

(1) 词典的规模太小。绝大部分词典的规模在一万词以下，无法很好的涵盖瞬息万变的 UGC 信息。

(2) 词典中的词太过正式。UGC 信息的特点是口语化，与词典中太过正式的词不符。情感词典应多涵盖一些网络词汇，例如，“进水”(贬义)、“给力”(褒义)等。

(3) 词典中仅包括词语，而没有词组。很多词语单独来看没有极性，然而，合并到一起就具有一定的情感倾向性，例如，“早知道”(贬义)、“怎么又”(贬义)等。

由于以上问题的存在，为情感分析的很多研究任务带来了困扰。海量的微博数据为扩大已有的情感词典规模提供了新的契机。在英文词典方面，Google 公司的研究者提出了一种基于图传播的算法，在网络上挖掘出 17 万余词的大规模的情感词典，在情感分类任务上取得了很好的效果[3]。Mohammad 等人从 tweets 中生成了一个 122 万词/词组的大规模的情感词典[4]。然而，在中文情感词典方面，还没有类似的大规模词典出现。基于此，本文提出了一套面向海量微博的大规模情感词典构建算法，并将其应用于情感分析的经典任务——情感分类上，实验证明了该词典的有效性。

2 相关研究

情感词语又称极性词、评价词，特指带有情感倾向性的词语。显然，情感词语在情感文本中处于举足轻重的地位，情感词语的抽取和极性判断在情感分析领域创建伊始就引起了人们极大的兴致。基于前人大量的研究工作，情感词语的抽取和判别主要分为基于语料库、基于词典及基于图模型三种方法。

基于语料库的方法主要是利用大语料库的统计特性，观察一些现象来挖掘语料库中的情感词语并判断极性，例如，由连词(如 and 或 but)连接的两个形容词的极性往往存在一定的关联性[5-7]。该方法最大的优点在于简单易行，缺点则在于可利用的评论语料库有限，同时情感词语在大语料库中的分布等现象并不容易归纳。

基于词典的方法主要是使用词典中的词语之间的词义联系来挖掘情感词语。这里的词典一般是指使用 WordNet 或 HowNet 等[8-9]。基于词典的方法的优点在于获取的情感词语的规模非常可观，但是由于很多词存在一词多义现象，构建的情感词典往往含有较多的歧义词。

基于图的方法主要将要分类的词语作为图上的点，利用词语之间的联系形成边来构建图，继而采用各种基于图的迭代算法来完成词语的分类[3,10]。基于图的方法是一种新颖的方法，它可以灵活地将词语间的各种联系作为特征融入图中，继而进行迭代计算。然而，寻找更为有效的词语间特征以及如何选取图算法是值得深入研究的问题。

大部分现有的情感词典规模都在一万词语以下，给情感分析的很多研究任务带来了困扰。为了解决情感词典的规模问题，在英文词典方面，Google 公司的研究者提出了一种基于图传播的算法在网络上挖掘出 17 万余词的大规模的情感词典[3]；此外，还有研究者从 tweets 中生成了一个大规模的情感词典，包含 62 468 词语、677 698 二元词组和 480 010 不连续的二元对[4]。这些词典均在英文情感分类任务上取得了很好的结果。然而，在中文情感词典方面，情感词典的规模依然很有限。例如，北京大学情感词典共有 449 个词，大连理工大学情感词典共包括27 466 个词[11]，清华大学情感词典共包括 10 036 个词[12]。基于此，本文致力于构建大规模的中文情感词典。

3 算法介绍

3.1 总体流程

面向微博领域的大规模情感词典的构建流程如图 1 所示。该情感词典构建算法包含两个步骤。

(1) 表情符种子获取：利用提前构建好的情感词语种子，在一个较小规模的微博语料上，为所有的表情符进行情感归类及重要性排序，从而为每类情感选择出一些相关性较高的、具有代表性的表情符。

(2) 情感词语/词组情感分值计算：利用上一步获得的表情符种子，在一个较大规模的微博语料上，为所有候选情感词语计算情感分值(本文使用

① http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html

unigram、bigram 和 trigram 作为候选情感词语），最后根据求得的所有候选情感词语的情感分值的符号与量级来构建情感词典。

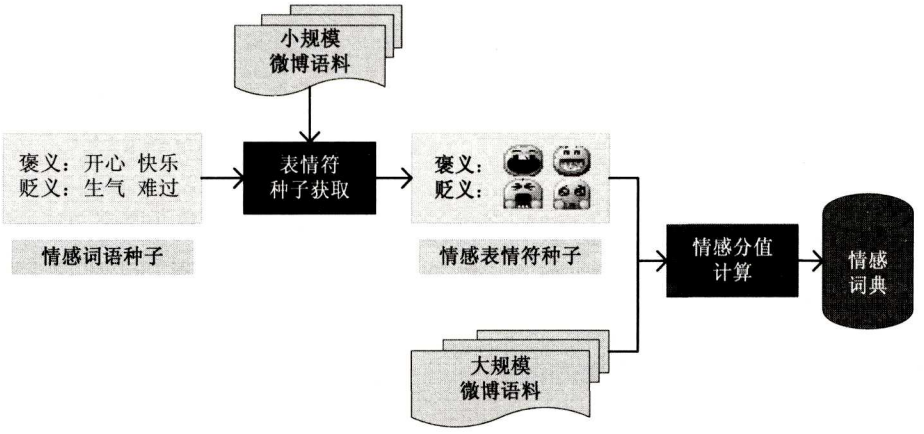

图 1　面向微博领域的大规模情感词典的构建流程

根据调研，有许多研究者直接使用情感词语做种子来构建情感词典。然而，种子情感词语的规模毕竟有限，其覆盖率较低，而表情符的覆盖率要远远超过种子情感词典，使用表情符做种子可以大大提高最终构建的情感词典的规模。因此，本文在前人工作的基础上又增加了一步，先利用情感词语种子来获取表情符种子，然后利用获取的表情符种子来构建情感词典。接下来就对本文使用的面向微博的大规模情感词典构建算法中包含的两个核心模块进行详细介绍。

3.2　表情符种子获取

在新浪微博中，用户可以使用的表情符的数量达上千个。然而，真正被用户所广泛使用且能够较准确代表用户情感的表情符却只有很少的一部分。因此，我们不仅需要对所有的表情符进行分类，以确定它们的情感倾向，而且需要对每种类别的表情符的重要性进行排序，以获取最能够代表对应情感倾向的表情符，移除情感倾向模糊的表情符。受 LI 等人工作的启发[13]，我们使用式(1)来衡量表情符 e_j 属于情感倾向 i($i\in$ {positive, negative})的概率。

$$P_i(e_j)=\frac{\sum_{k=0}^{k=n}\mathrm{COF}(e_j,w_k)}{\sum_{i=0}^{i=m}\sum_{k=0}^{k=n}\mathrm{COF}(e_j,w_k)}\times\log(\mathrm{freq}(e_j)) \tag{1}$$

其中，$\mathrm{COF}(e_j,w_k)$ 表示表情符 e_j 与情感倾向 i 的词典(有两个种子词典：褒义种子词典和贬义种子词典)中第 k 个情感词 w_k 在微博语料中的共现频次；$\mathrm{freq}(e_j)$ 代表表情符 e_j 在语料中出现的频次；分母的作用是归一化。参数 n 代表了情感倾向 i 的词典的种子词语的个数，m 代表了情感倾向词典的个数，在本文中，$m=2$。

为了统计 w_k 的频次和 e_j 与 w_k 的共现频次，我们需要爬取一定数量的微博语料。本文使用微博爬虫通过新浪微博提供的 API 接口爬取到的 2013 年 3 月的数据，经过过滤、去除广告等预处理后，得到约 400 万条的微博数据。本文用到的种子情感词典来自于北京大学发布的情感资源，共包括褒义词 86 个，贬义词 396 个，褒义表情符 11 个，贬义表情符 17 个。

基于这些微博语料和词典资源，我们可以为上千个微博表情符计算他们属于褒义倾向和贬义倾向的概率，并根据概率值进行排序。我们人工挑选了一些排序靠前的表情符来代表对应的情感倾向，具体如表 1 所示。

表 1　表情符及其情感倾向

情感倾向	表　情　符
褒义	([哈哈]) ([嘻嘻]) ([偷笑]) ([花心]) ([得意]) ([高兴]) ([蛋糕]) ([开心]) ([欢欣鼓舞]) ([太开心])
贬义	([伤心]) ([委屈]) ([悲伤]) ([失望]) ([可怜]) ([抓狂]) ([愤怒]) ([泪]) ([上火]) ([怒气]) ([怒骂]) ([怒]) ([吃惊]) ([惊恐]) ([惊讶])

3.3 情感词语/词组情感分值计算

为了能够获取更大规模的情感词典，单使用表情符来代替情感词语做种子是远远不够的，更重要的是必须有一个大规模的语料集。为此，本文整合了通过微博爬虫爬取的从2013年4月到2014年3月共12个月的微博数据作为生成情感词典的语料，约14.6亿条微博数据，大小约360GB。获取情感词典的种子使用的则是我们前文所述的已经构建好的种子表情符。

我们假设微博中使用的表情符的情感倾向和微博文本本身的情感倾向是一致的，那么如果一条微博中包含任意一个褒义的表情符种子，那么我们就认为这条微博是褒义的；同样的，如果一条微博中包含任意一个贬义的表情符种子，那么我们就认为这条微博是贬义的；如果一条微博中同时包含褒义和贬义的表情符种子，那么我们就弃用该微博。根据这种方法，我们共收集了约4GB的褒义微博数据，约2.5GB的贬义微博数据。后文将在这个数据集的基础上来生成情感词典。

我们从上述微博语料中抽取N-gram(unigram、bigram和trigram)作为候选情感词，目的即为候选情感词进行情感分值的计算以从候选情感词中抽取出真正的具有情感倾向性的词语。在计算N-gram的情感分值的过程中，我们主要使用到了点互信息。点互信息(pointwise mutual information，PMI)，常用于衡量两个变量 x 和 y 之间的相关性。点互信息的数学表达如式(2)所示。

$$\mathrm{PMI}(x,y) = \log \frac{p(y \mid x)}{p(y)} \tag{2}$$

每一个候选情感词 w 与对应情感倾向的相关性分值可以使用式(3)计算得到。

$$\mathrm{SCORE}(w) = \mathrm{PMI}(w, E_p) - \mathrm{PMI}(w, E_n) \tag{3}$$

其中，w 为候选情感词，E_p 代表所有情感极性为褒义的表情符种子，E_n 代表所有情感极性为贬义的表情符种子。SCORE的极性表示 w 与哪种情感类别相关，具体的，正值为褒义，负值为贬义；SCORE的量级则表示 w 与相应情感类别相关的程度，具体的，值越大，越相关。

基于以上简单的文本统计算法，本文得到了一个较大规模的情感词典。该词典的详细信息见表2。表中数据显示，该方法获取的情感词典的规模非常大，在后续实验中，我们发现该词典的大规模为计算带来了较大困难。此外，SCORE绝对值比较小的一些候选情感词很多不具有情感倾向性。因此，在实验部分，我们将通过设置阈值来对该词典进行过滤，以获取一个有效的大规模情感词典。

表2 过滤前的大规模情感词典

情感词类型	褒义/个	贬义/个
unigram	843 546	340 059
bigram	962 063	996 860
trigram	993 180	993 871

4 实验与分析

为了验证本文构建的大规模情感词典的有效性，我们将其用在情感分析的经典任务——情感分类上面。具体为，判断一条微博的情感倾向为褒义、贬义还是中性。

4.1 数据

为了反映微博数据的真实情况，我们从新浪微博的数据中随机选择了8 512条微博，并请三位标注者同时对这些微博数据进行了人工标注。标注者根据微博文本的情感倾向将所有微博分为褒义、贬义和中性三个类别，若有标注结果不一致的情况，则使用投票的方法决定微博文本的情感类别。这些微博数据的人工标注结果统计如表3所示。

表3 人工标注结果统计

褒义	贬义	中性	总和
2 401	2 853	3 258	8 512

鉴于语料规模有限，我们使用五折交叉验证的方法来进行验证，实验使用支持向量机模型，对语料进行褒义、贬义或中性的三元情感分类。

4.2 词典规模有效性验证

由表2可知，本文构建的情感词典的规模是非常庞大的。为了获取有效的情感词典，我们将其应用于情感分类任务上，并选择了简单有效的基于特征分类的情感分类算法[14]。具体的，针对一条微博，提取的特征除了BOW(bag of words)特征外，根据大规模词典还加入了两维特征，分别是该微博中包含词典中的褒义词的个数与贬义词的个数。由于本文构建的情感词典由unigram、bigram和

trigram 构成，我们分别就这三个部分对情感分类性能的影响进行了实验，以确定词典中这三个部分的规模。

我们依次累积增大三个词典的使用规模，以验证词典规模对情感分类性能的影响。图 2、图 3 及图 4 分别展示了情感分类的准确率随 unigram、best unigram + bigram[①] 和 best unigram + best bigram + trigram[②] 情感词典规模的变化而变化的趋势。其中，纵轴表示微博情感分类的准确率，横轴表示使用的情感词典的规模（例如，BOW+2 000 表示除了使用 BOW 特征以外，还使用褒义、贬义各 2 000 个词语，即总共 4 000 个情感词语）。

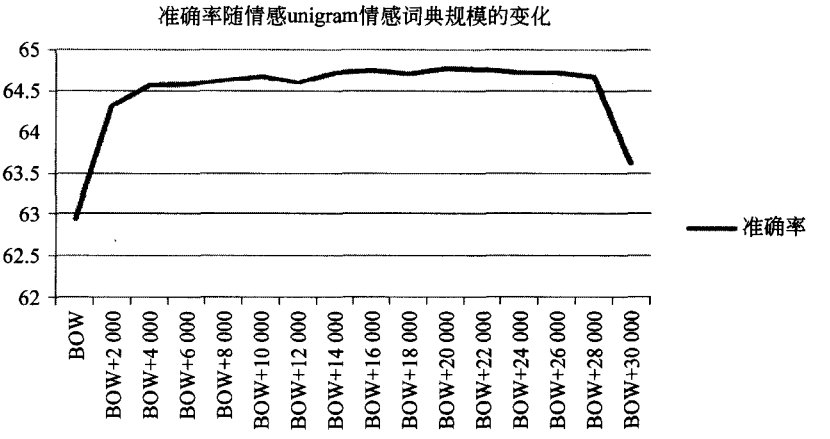

图 2　情感分类的性能与 unigram 情感词典规模的相关性

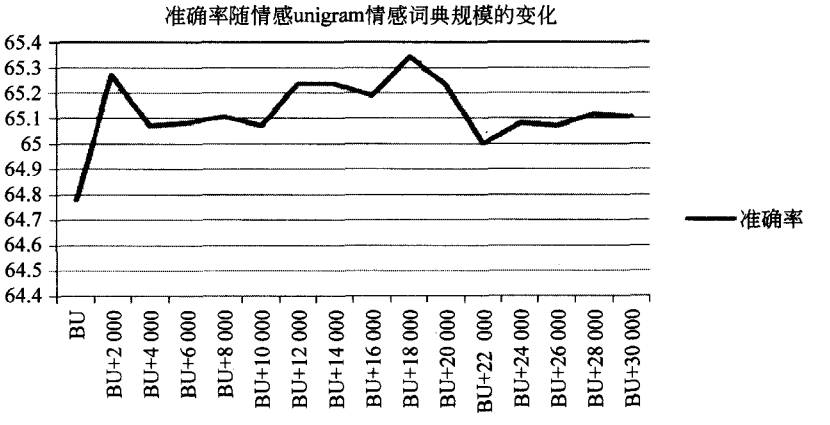

图 3　情感分类的性能与 bigram 情感词典规模的相关性

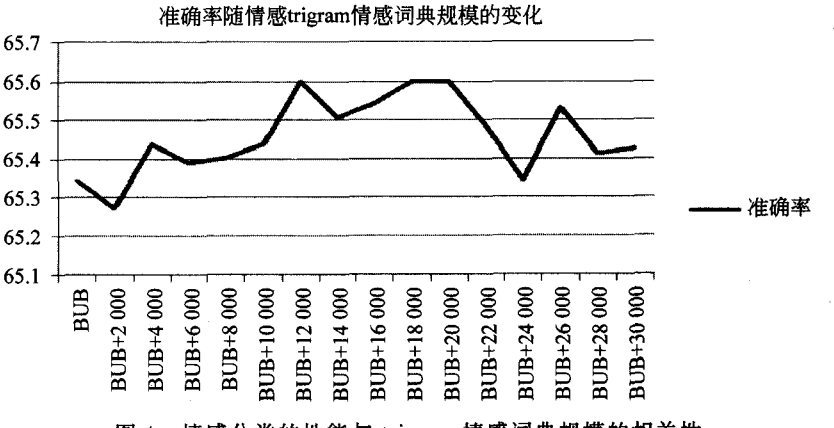

图 4　情感分类的性能与 trigram 情感词典规模的相关性

① best unigram 是指我们在测试 bigram 情感词典规模的时候，是在最优的 unigram 的基础上进行测试的。

② best unigram+best bigram 是指我们在测试 trigram 情感词典规模的时候，是在最优的 unigram 和 bigram 的基础上进行测试的。

通过分析可知：(1)对于 unigram 情感词典，随着对其使用规模的增加，情感分类的准确率有显著提升；当其褒义、贬义情感词语各使用 20 000 时分类准确率达到最大，比单独使用 BOW 提高了 1.84%；当其褒义、贬义情感词各使用 28 000 时性能开始急速下降，这可能与后续加入的 unigram 的情感分值较低，以至于引入大量噪声有关。(2)对于 bigram 情感词典，当其褒义、贬义情感词各使用 18 000 时性能达到最优，在 BOW＋unigram 获取的最优性能的基础上又进一步提高了 0.56%；当其褒义、贬义情感词语各使用 22 000 时性能明显下降；进一步加大使用规模后虽然性能稍有提升，但普遍较低。(3)对于 trigram 情感词典，当其褒义、贬义情感词各使用 12 000 时性能达到最优，在 BOW＋unigram＋bigram 获取的最优性能的基础上又进一步提高了 0.26%；进一步加大使用规模后的分类性能也同 bigram 一样，虽然局部有所提升，但准确率普遍较低，我们认为这与新加入的情感词语的质量较差有很大关系。

基于以上的分析，最终我们的大规模情感词典的分布情况如表 4 所示。

表 4 过滤后的大规模情感词典

情感词类型	褒义/个	贬义/个
unigram	20 000	20 000
bigram	18 000	18 000
trigram	12 000	12 000
共计	50 000	50 000

4.3 与其他词典的对比

除了我们自己构建的面向微博的大规模情感词典以外，本文还使用了其他四个开源的情感词典资源，它们分别来自清华、北大、大连理工及知网，详见表 5。为了对比我们自己构建的情感词典与其他的情感词典资源的性能，本文进行了如下的实验，其中：

表 5 词典规模统计

词典	褒义	贬义	总和
清华(Tsinghua)	5 567	4 468	10 035
北大(Peking)	95	420	515
大连理工(DUT)	11 043	10 646	21 689
HowNet	4 528	4 320	8 848
我们的词典(Our)	50 000	50 000	100 000

(1) BOW(bag of words)＋ALL(all lexicon feature)：表示在 BAG OF WORDS 特征的基础上使用全部的情感词典资源(包括我们自己构建的词典以及其他的四个词典资源)；

(2) BOW ＋ ALL-Our：表示在 BAG OF WORDS 特征的基础上使用除了我们自己构建的情感词典以外的全部词典资源；

(3) BOW＋ALL-HowNet：表示在 BAG OF WORDS 特征的基础上使用除了知网的情感词典以外的全部词典资源；

(4) BOW ＋ ALL-DUT：表示在 BAG OF WORDS 特征的基础上使用除了大连理工的情感词典以外的全部词典资源；

(5) BOW ＋ ALL-Peking：表示在 BAG OF WORDS 特征的基础上使用除了北大的情感词典以外的全部词典资源；

(6) BOW＋ALL-Tsinghua：表示在 BAG OF WORDS 特征的基础上使用除了清华的情感词典以外的全部词典资源。

各词典的性能对比详见表 6。

表 6 各情感词典性能对比

系统	Accuracy/%	Change/%
BOW＋ALL	65.73	—
BOW＋ALL-Our	64.00	−1.73
BOW＋ALL-HowNet	65.47	−0.26
BOW＋ALL-DUT	65.58	−0.15
BOW＋ALL-Peking	65.60	−0.13
BOW＋ALL-Tsinghua	65.67	−0.06

通过分析表 6 可知，本文构建的情感词典的性能要显著得优于其它四类情感词典。但是从表 6 也可以发现，本文的情感词典并不能够完全替代其他四类情感词典。在使用了本文构建的情感词典的基础上再使用这些情感词典资源，对情感分类的性能仍能有一定的提升。

4.4 在情感分类任务上的应用

本文选取了情感分类任务作为构建的大规模情感词典的应用点。为了测试该词典的有效性，我们借鉴最经典的英文情感分类系统 NRC-Canada[10] 来构建我们的基于 SVM 分类器的情感分类系统。对比实验设计如下：

(1) Baseline：使用 NRC-Canada 系统中适用

于中文微博的特征，其中，在词典部分，我们使用了表 6 中除了我们的词典的所有四个词典作为特征进行微博情感分类；

(2) Baseline + Our(我们的词典)：在 Baseline 系统的基础上，引入两维特征，分别是该微博中包含词典中的褒义词的个数与贬义词的个数，基于此进行微博情感分类。

对比实验结果如表 7 所示。

表 7　在情感分类任务上的对比实验

系统	Accuracy/%
Baseline	64.74
Baseline+Our	65.87

通过分析表 7 可知，本文构建的面向微博的大规模情感词典能够显著得提升微博情感分类的性能(1.13%)，充分证明了该词典的有效性。

5　结论

为了解决现有的中文情感词典的规模小、口语化词语少以及缺少情感词组等问题，本文面向海量的微博数据，提出了一种简单的构建大规模情感词典的方法，并构建了一个规模为 10 万词语/词组的情感词典。本文将该词典应用于情感分类任务上，实验结果表明：本文构建的大规模情感词典的性能要远超其他中文情感词典；此外，将本文的情感词典融入经典的微博情感分类算法中，能够显著的提高该算法的实验性能。

参考文献

[1] 赵妍妍，秦兵，刘挺. 文本情感分析[J]. 软件学报，2010，21(8)：1834-1848.

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

[3] L Velikovich, S Blair-Goldensohn, K. Hannan, R McDonald. The viability of web-derived polarity lexicons[C]// Proceedings of the NAACL, 2010: 777-785.

[4] S Mohammad, S Kiritchenko, X Zhu. NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets[C]//Proceedings of the Second Joint Conference on Lexical and Computational Semantics (* SEM), 2013: 321-327.

[5] V Hatzivassiloglou, K McKeown. Predicting the semantic orientation of adjectives [C]//Proceedings of the EACL, 1997: 174-181.

[6] J Wiebe. Learning subjective adjectives from corpora [C]//Proceedings of the AAAI, 2000: 735-740.

[7] P Turney, M Littman. Measuring praise and criticism: Inference of semantic orientation from association[J]. ACM Trans. on Information Systems, 2003, 21(4): 315-346.

[8] SKim, E Hovy. Automatic detection of opinion bearing words and sentences[C]//Proceedings of the IJCNLP, 2005: 61-66.

[9] S Kim, E Hovy. Identifying and analyzing judgment opinions[C]//Proceedings of the NAACL, 2006: 200-207.

[10] D Rao, D Ravichandran. Semi-Supervised polarity lexicon induction [C]//Proceedings of the EACL, 2009: 675-682.

[11] 徐琳宏，林鸿飞，潘宇，等. 情感词汇本体的构造[J]. 情报学报，2008，27(2)：180-185.

[12] 李军. 中文评论的褒贬义分类实验研究[D]. 清华大学硕士学位论文，2008.

[13] F Li, S Pan, O Jin, et al. Cross-Domain Co-Extraction of Sentiment and Topic Lexicons[C]//Proceedings of the 50th ACL, 2012: 410-419.

[14] B Pang, L Lillian, V Shivakumar. Thumbs up? Sentiment Classification using Machine Learning Techniques [C]//Proceedings of the EMNLP, 2002: 79-86.

赵妍妍(1983—)，讲师，主要研究领域为情感分析。

E-mail: yyzhao@ir.hit.edu.cn

石秋慧(1989—)，硕士研究生，主要研究领域为情感分析。

E-mail: qhshi@ir.hit.edu.cn

秦兵(1968—)，教授，主要研究领域为文本挖掘、情感分析。

E-mail: bqin@ir.hit.edu.cn