一种基于特征簇的微博短文本情感分类方法 *

周咏梅[1,2+]，王伟[1]，阳爱民[1,2]，林江豪[3]，方泽锋[1]

[1](广东外语外贸大学思科信息学院，广东广州 510006)

[2](广东外语外贸大学语言工程与计算实验室，广东广州 510006)

[3](广东外语外贸大学 财务处，广东广州 510420)

A Method for Sentiment Classification of Microblog Short Text Based on Feature Clusters*

ZHOU Yong-mei[1,2], WANG Wei[1], YANG Ai-min[1,2+], LIN Jiang-hao[3], Fang Ze-feng[1]

[1](Cisco School of Informatics, Guangdong University of Foreign Studies, Guangzhou 510006, Guangdong)

[2](Laboratory for Language Engineering and Computing, Guangdong University of Foreign Studies, Guangzhou 510006)

[3](Financial Department, Guangdong University of Foreign Studies, Guangzhou 510420, China)

+ Corresponding author: E-mail: yongmeizhou@163.com

Abstract: For the data sparseness of Microblog short text caused by the great scale of all features, little features of one single text and so on, we propose a method of sentiment classification of Microblog text based on feature clusters. Firstly, the word2vec model is used to learn the latent semantic relations between words from a large-scale corpus, and then each word is represented with a multi-dimensional vector. Secondly, the affective features, which are extracted with the sentiment lexicon, are merge into feature clusters based on the method of computing the word similarity with the term vector, and then these feature clusters are used construct the text vector with low-dimension. Lastly, the machine-learning algorithm is used to construct the classification of Microblog short text. In the experiment, the method we presented is feasible and effective in reducing the dimensionality of the affective features and shows effectiveness on text sentiment classification.

Key words: microblog sentiment; data sparseness; term vector; feature cluster; machine learning

摘　要: 针对由微博短文本特征规模大、自身特征较少等特点导致的数据稀疏性，提出一种基于特征簇的微博情感分类方法。提出的分类方法以大规模语料库为基础，利用 *word2vec* 模型学习词语之间潜在的语义关联，将单个词语表示成多维向量的形式；结合情感词典，提取出微博文本的情感特征集，在基于词向量计算词语相似度方法的基础上，将情感特征合并为特征簇，以此构造低维的文本向量；最后利用机器学习算法，构建情感分类器，实现微博短文本的情感分类。实验结果表明，本文提出的方法对情感特征的降维是可行和有效的，并且取得很好的情感分类效果。

关键词: 微博情感；数据稀疏；词向量；特征簇；机器学习

中图法分类号: TP391　　**文献标识码**: A

*The National Social Science Funding Project of China under Grant No. 12BYY045 (国家社会科学基金项目); the New Century Excellent Talents Foundation from Ministry of Education of China under Grant No. NCET-12-0939 (教育部新世纪优秀人才支持计划); the Technology Innovation Project of Department of Education of Guangdong Province under Grant NO. 2013KJCX0067(广东省教育厅科技创新项目); the Postgraduate Research & Innovation Project of Guangdong University of Foreign Studies under Grant No. 14GWCXXM－36 (广东外语外贸大学研究生科研创新项目); the School Project of Regular institutions of higher learning in Guangdong University of Foreign Studies under Grant No. 14Q3(广东外语外贸大学校级项目); the Innovation & Entrepreneurship Training Program Project in Guangdong University of Foreign Studies under Grant No. 201511846021(广东外语外贸大学创新创业训练计划项目).

作者简介：周咏梅(1971-),女,教授,主要研究领域为文本情感分析,微博情感分析,E-mail:yongmeizhou@163.com;王伟(1991-),男 ,硕士研究生,主要研究领域为文本情感分析;阳爱民(1970-),男,博士,教授,主要研究领域为文本情感分析,机器学习,模式分类,微博情感分析;林江豪(1985-),男,硕士,主要研究领域为文本情感分析;方泽锋(1994-),男,本科,主要研究领域为文本情感分析.

1 引言

互联网的蓬勃发展，在线文本呈现爆炸式的增长。特别是微博和移动互联网等社交平台与技术的普及，方便了网民观点的表达与传播，导致产生了大量主观性的网络文本信息。针对主观性文本中所包含的用户观点、情感和情绪等进行分析与挖掘的研究，已成为自然语言处理技术的一个热点方向，并且被成功应用到舆情监测、产品营销和股价预测等领域，具有极大的实用价值和应用前景。微博文本往往以短文本形式存在，而且与传统的长文本相比，存在用词比较随意、文本格式较不规范和表意方式多样等特点，以及由于文本字数较少引起的天然极稀疏性导致很难提取出有效的内容特征。这些特点对微博文本的信息挖掘带来了很大的挑战[1]。

目前，在微博文本情感分析领域，梁军等[2]人则尝试利用深度学习对中文微博进行情感分析，利用递归神经网络发现与任务相关的特征，并根据句子词语间前后的关联性引入情感极性转移模型加强对文本关联性的捕获。杨佳能等[3]人提出一种基于语义分析的中文微博情感分类方法，引入情感词典、网络词典和表情符号，对微博文本进行依存句法分析，以此构建情感表达式树，最后通过定义规则计算微博文本的情感强度和情感倾向。贺飞艳等[4]人结合 *TF-IDF* 与方差统计方法，提出一种细粒度情感特征抽取方法。文献[5-6]则利用微博中的表情图片和情绪图片构建微博情感语料库，在此基础上构建分类器，实现微博文本情感的分类。针对微博情感特征稀疏性问题，许多学者结合微博文本的特点提出了不同的分析方法。吴方照等[7]人结合社会科学的相关理论，尝试利用社交语境帮助解决情感分析所面临的稀疏性强和噪声大的困难，提出将 *Lasso* 方法加进 *logistic regression* 模型中以提高模型的鲁棒性。周剑锋等[8]人定义了 7 种词语词性搭配模型，以微博语料为基础，构建二元词语搭配词库，提出情感特征权值的计算方法 *PMI-IR-P*。*Yuki Yamamoto* 等[9]人利用微博文本中表情符这类元数据，根据其不同作用，将其划分为“强调”，“削弱”，“转换”和“添加”四中角色，提出一种基于表情符角色对 *Twitter* 文本进行情感分类的方法。*Farhan Hassan Khan* 等[10]人提出一种融合表情符、*SentiWordNet* 和种子情感词的 *Twitter* 文本意见挖掘框架，并取得了较高的准确率。王磊等[11]人则将“主题”这个概念引入到短文本的分析中，利用 *LDA* 主题模型发掘文本的潜在主题特征，并将其融入到情感模型中，实验证明主题模型在情感分类任务上有着良好的表现。

现有研究针对短文本特征的稀疏性进行了多方面的尝试。文献[12]经过统计得到，短文本平均文本长度为 70.41114 词，文本稀疏度达 0.16573%，并指出处理短文本的困难主要存在两个方面，一方面是特征属性规模庞大，另一方面是单一文本自身特征太少。鉴于此，本文从降低微博文本特征维数方面着手，提出一种基于特征簇的微博短文本情感分类方法。这种方法以大规模语料库为基础，使用 *word2vec* 模型训练语料，学习词语之间的语义关系，并用词向量形式表示词语；然后，通过构建情感词典，提取出微博文本的情感特征集，利用基于词向量计算词语相似度的方式，将情感特征合并为特征簇，以此构造维数较低的文本向量；最后，利用支持向量机（*Support Vector Machine, SVM*）方法实现微博文本情感的分类，实验表明了本文提出方法的有效性。

本文的第 2 部分介绍提出的微博短文本情感分类方法的概述，第 3 部分介绍情感特征的降维，第 4 部分进行实验以及分析结果，第 5 部分是结论和将来的研究工作。

2 基于特征簇的微博短文本情感分类方法方法概述

本文提出的基于特征簇的微博短文本情感分类方法基本框架如图 1 所示。主要包括两部分，词向量模型的学习和特征簇的构造。

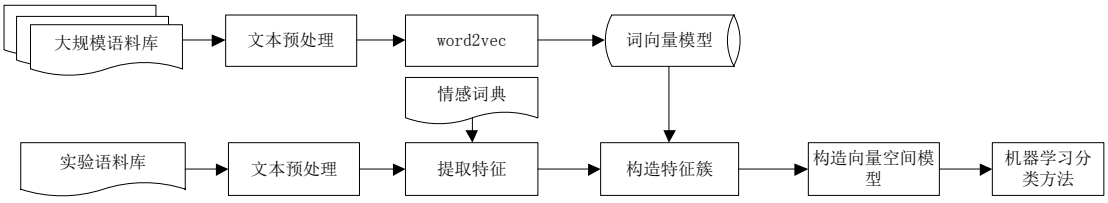

Fig. 1 The framework of sentiment analysis for Microblog short texts

图1 提出的微博短文本情感分类方法基本框架

本文通过设计爬虫程序从微博平台采集微博数据，以此构建大规模微博语料库，并使用分词工具对语料进行分词、词性标注，过滤停用词和无效字符，并统计词频；利用 *word2vec* 模型对预处理过的微博语料进行训练，学习词语之间潜在的语义关联，得到词向量模型，用多维向量的形式表示单个词语，即 $w_t=<V_1,V_2,\cdots,V_n>$，V_i 代表特征项。

通过结合情感词典，提取出微博文本中所有的情感特征，利用训练好的词向量模型衡量特征之间的语义相似性，设计算法将符合要求的情感特征合并为特征簇，把包含多个情感特征的单个特征簇作为文本向量的新特征项，构造向量空间模型，与机器学习的分类方法结合，实现微博短文本的情感分类。

其中，*word2vec*[13]是一款利用深度学习的思想，通过训练模型将单个词语表征为实数值向量的工具。其核心架构主要基于两个模型：*CBOW*（*Continuous Bags-of-Words*）和 *Skip-gram* 模型，如图 2 所示。*CBOW* 模型由“输入层—映射层—输出层”三层结构组成，模型原理与神经网络概率语言模型（*Neural Netword Language Model, NNLM*）类似，都是利用当前词 w_t 的上下文 $w_{t-k},\cdots,w_{t-1},w_{t+1},\cdots w_{t+k}$ 预测当前词 w_t 的概率 $p(w_t\mid w_{t-k},\cdots,w_{t-1},w_{t+1},\cdots w_{t+k})$。与 *NNLM* 不同的是，*CBOW* 模型为了提高计算效率，去掉了最耗时的非线性隐藏层，并让输入层的所有词语共享映射层。如图 2 左边模型所示，*CBOW* 模型使用单词 w_t 周边的词语作为输入，在映射层做加权处理，然后输出单词 w_t。*Skip-gram* 模型也是三层结构，但是处理过程与 *CBOW* 模型正好相反，它是利用当前词 w_t 预测了上下文的概率 $p(w_i\mid w_t), t-k\le i\le t+k$，如图 2 右边模型所示。一般地，*Skip-gra*m 模型理解为根据当前词预测上下文构成的语境，而 *CBOW* 模型是根据语境预测目标词。

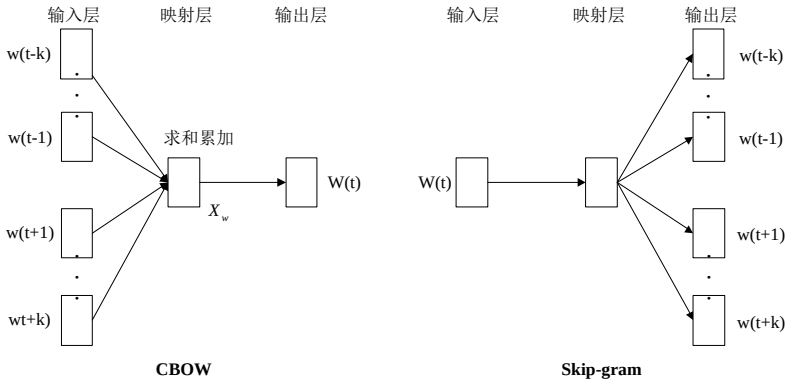

Fig. 2 The structure of Continuous Bag of Words and Skip-gram

图2 CBOW模型和Skip-gram模型的框架

word2vec 考虑了当前词的上下文信息，由此学习到的词向量，包含了丰富的语义和语法关系。词向量之间的关系，通过加减代数运算可以被体现出来，例如，对“国王”，“男人”，“女王”和“女人”四个词语的词向量进行向量运算，会得到 $V(\text{"国王"})-V(\text{"女王"})\approx V(\text{"男人"})-V(\text{"女人"})$[14]。另外，两个词向量的余弦值可以用来衡量两个词语之间的相似程度。例如，表 1 是利用 *word2vec* 对语料进行训练之后，得到的分别与“高兴”、“烦恼”最接近的词语。训练所使用的语料是从新浪、腾讯微博收集的共 50 万条微博文本。

Table 1 Similar word

表 1 相似词

高兴		烦恼	
相似词语	相似度	相似词语	相似度
荣幸	0.5104711	困难	0.2797439
激动	0.5077044	痛苦	0.27539194
兴奋	0.49424824	困扰	0.26958287
欣慰	0.4543456	难题	0.2676941
自豪	0.45084906	困惑	0.2671121
感激	0.402286	困境	0.26432782
开心	0.39744002	麻烦	0.256484

3 基于特征簇的情感特征构造

3.1 提取情感特征

本文只对包含情感词的微博文本进行情感极性判断，对不包含情感词的微博文本暂不考虑。利用情感词典匹配出微博文档集中的情感词，将情感词作为情感特征构造向量来表示文本。

提取情感特征前，需要构造情感词典。一个较完整的情感词典对提取微博文本中的重要情感信息很重要。本文把 *HowNet* 极性词典、台湾大学的 *NTUSD* 情感词典和情感本体库 [15]合并后，把重复词语去掉，得到较完整的情感词集。利用每个词语在不同词典的标注结果，对其情感倾向进行褒贬投票。如果投票倾向一致，则将情感词自动加入本文所用情感词典，否则采用人工方式对其进行标注并且多次校对。

得到情感特征后，将文本表示成向量的形式，如式（1）所示，其中，V_i代表特征项，n 代表特征的数量：

$$\vec{d} = < V_1, V_2, \ldots, V_i, \ldots, V_n > \tag{1}$$

3.2 基于特征簇的文本向量表示

由 3.1 节得到的文本向量维数一般较大，特征项的数量往往比一个微博文本 d_i所包含特征的数量多很多，使得文本向量中相当数量元素的值为零。因此，需要对情感特征进行降维，规避数据稀疏性的问题。

本文研究目的限定在对微博文本进行情感极性的判定，即把微博的情感分类为正向、负向和中性，对于微博表达的情感强烈程度不进行讨论。以正向情感为例，对微博文本①和②分别提取情感特征“好看”和“合身”，可以得出微博①表达的情感明显比微博②强烈，但是两条文本体现的情感倾向是一致的。另外，将微博②的情感特征“合身”替换为“好看”，得到微博文本③，文本原先的情感倾向并不会改变，情感更加强烈，反而对微博的情感分类更有帮助。

①今天的电影很好看！

②刚买的衣服穿上去合身。

③刚买的衣服穿上去好看。

而且对情感特征进行词频统计，得到特征词“合身”词频较低，容易导致数据稀疏性。所以构造文本向量时，可以将这两个特征词进行合并为单个的特征项，即<好看，合身>，起到降低文本向量维数的效果。因此，本文提出将情感特征合并成特征簇 $C_i=[V_1, V_2, \ldots, V_t]_i$，$t$ 为特征簇中特征词的数目，实现情感特征降维的目的。文本向量用特征簇表示如式（2）所示，其中，C_k代表特征簇，m 代表特征簇的数目。

$$\vec{d} = < C_1, C_2, \ldots, C_i, \ldots, C_m > \tag{2}$$

构造特征簇是一个将由情感特征构成的高维文本向量转化为由特征簇表示的低维文本向量的过程，以图 3 为例。

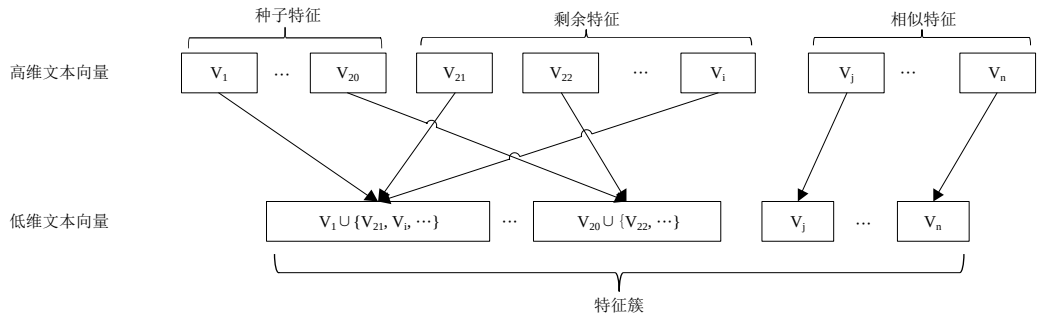

Fig. 3 Constructing vector clusters

图3 构造特征簇

文本将构造特征簇过程中的情感特征划分为三类：种子特征、相似特征和剩余特征，分别对其定义：

①种子特征：利用式（3）计算所有情感特征的卡方统计值 χ^2，对结果进行排序，分别筛选出正负向情感特征中前 h 个 χ^2 值较大的特征项为种子特征，记为 $seedV_1, seedV_2, \cdots, seedV_{2h}$；

$$\chi^2(V_i, P_j) = \frac{(AD - BC)^2}{(A+B)(C+D)} \tag{3}$$

式（3）中，P_i 表示情感类别，A 表示包含情感特征 V_i 且属于 P_i 类的微博文本数；B 表示包含情感特征 V_i 但是不属于 P_i 类的微博文本数；C 表示属于 P_i 类但是不包含情感特征 V_i 的微博文本数；D 表示既不属于 P_i 类也不包含 V_i 的微博文本数。

②相似特征：使用 *word2vec* 学习到的词向量模型，计算出种子特征的 *topN* 个相似词，将其与情感特征进行匹配，将匹配到的相似词作为相似特征，记为 $simV_i$；

③剩余特征：情感特征集中除了种子特征和相似特征以外的特征项，定义为剩余特征，记为 $surV_i$；

3.3 特征簇构造算法

本文提出的特征簇构造具体算法过程如算法 1 所示。模型输出得到特征簇 $\{C_1, C_2, \ldots, C_m\}$，通过式（2）的形式表示文本向量，作为机器学习方法的输入，构造本文的微博情感分类器，实现微博短文本情感分类的任务。

算法 1：特征簇构造算法

输入：情感特征 $\{V_1, V_2, \ldots, V_n\}$

输出：特征簇 $\{C_1, C_2, \ldots, C_m\}$

步骤 1：计算所有情感特征 V_i 的卡方统计值，$x\_array[n] = \{\chi^2(V_1), \chi^2(V_2), \ldots, \chi^2(V_n)\}$，对结果进行排序 $sort(x\_array[n])$，筛选出 h 个正向种子特征和 h 个负向种子特征 $\{seedV_1, seedV_2, \ldots, seedV_{2*h}\}$，本文取 h 的缺省值为 10；

步骤 2：利用 word2vec 训练的词向量模型 v_Model，计算出每一个种子特征的 *topN* 个相似词，将得到的所有相似词合并去重，得到 $\{simW_1, simW_2, \ldots, simW_p\}$；将得到的相似词与情感特征进行匹配，即由 $\{simW_1, simW_2, \ldots, simW_p\} \cap \{V_1, V_2, \ldots, V_n\}$ 计算得到的特征项作为相似特征 $simV_i$，每一个相似特征被单独作为一个特征簇 $simC_k$；

步骤 3：计算 $\{V_1, V_2, \ldots, V_n\} - \left\{\{simW_1, simW_2, \ldots, simW_p\} \cap \{seedV_1, seedV_2, \ldots, seedV_{20}\}\right\}$ 得到剩余特征 $surV_j$；

循环 1：遍历每所有剩余特征；

步骤 4：利用词向量模型 v_Model，分别计算剩余特征 $surV_j$ 与正负向种子特征相似度之和，由 $p = \sum_{i=1}^{l=10} sim(surV_j, seedV_i^+) - \sum_{i=1}^{l=10} sim(surV_j, seedV_i^-)$ 确定剩余特征 $surV_j$ 的正负倾向；

步骤 5：计算出属于倾向 p 并且与剩余特征之间相似度最大的种子特征 $seedV_i$，将 $surV_j$ 与 $seedV_i$ 合并成到一个特征簇 $seedC_k, 1 \le k \le 20$；

结束循环 1；

步骤 6：将 $simC_k$ 和 $seedC_k$ 组合得到情感特征合并后的特征簇 $\{C_1, C_2, \ldots, C_m\}$；

结束：输出 $\{C_1, C_2, \ldots, C_m\}$。

4 实验及结果分析

4.1 实验数据

本文设计爬虫程序从新浪微博和腾讯微博平台共采集了 50 万条微博，构建语料库用于 *word2vec* 模型

的训练。情感分类实验数据采用 *NLP&CC2013*[1]和 *NLP&CC2014* 评测会议中微博情感分类任务所使用的语料，随机选取部分标注语料作为本文实验训练集和测试集，如表 2 所示。实验中 *word2vec* 参数的设置情况如表 3 所示。数据预处理采用中科院 *ICTCLAS* 分词工具对微博语料进行分词、词性标注。实验中的机器学习分类器选用 *SVM*，工具选取台湾大学林智仁开发的 *LibSVM*。

Table 2 Experimental data

表 2 实验数据

数据集	负向微博	正向微博	总数
训练数据集	1200	1200	2400
测试数据集	600	600	1200

Table 3 Parameter settings of word2vec

表 3 word2vec 参数设置

超参数	含义	取值
-size	向量维数	50~300，间隔 50
-window	上下文窗口大小	5
-sample	高频词亚采样的阈值	le^{-3}
-isCbow	使用 *cbow* 算法/*skip-gram* 算法	*true/false*

4.2 评测标准

本文对不包含情感词的微博文本暂不考虑，并且认为包含情感词的微博文本具有单一情感极性，分类结果只有正向或负向。对于每一个微博文本都能进行分类的语料集，评判分类器性能的准确率（*Precision*）、召回率（*Recall*）和 *F* 值是相等的。因此采用总体准确率作为本文方法的分类性能评价指标公式为：

$$Overall-accuracy=\frac{\sum_{c_i\in C} Correct(c_i)}{\sum_{c_i\in C} Doc(c_i)} \tag{4}$$

其中，*Over-accuracy* 代表总体准确率，$Correct(c_i)$是分类为 c_i 并且正确的微博文本数，$Doc(c_i)$是类别为 c_i 的微博文本总数。

4.3 实验结果及分析

本文实验将微博的情感分为正向和负向。设置两组实验验证本文方法的有效性，一组是实验参数设置不同值的情况下，分析情感分类准确率的变化；另一组是利用快速主成分分析法 *fastPCA*[16]对情感特征进行降维，与本文方法进行实验对比分析。

（1）实验参数对分类准确率的影响实验

1 自然语言处理与中文计算会议(*Natural Language Processing & Chinese Computing, NLP&CC*)

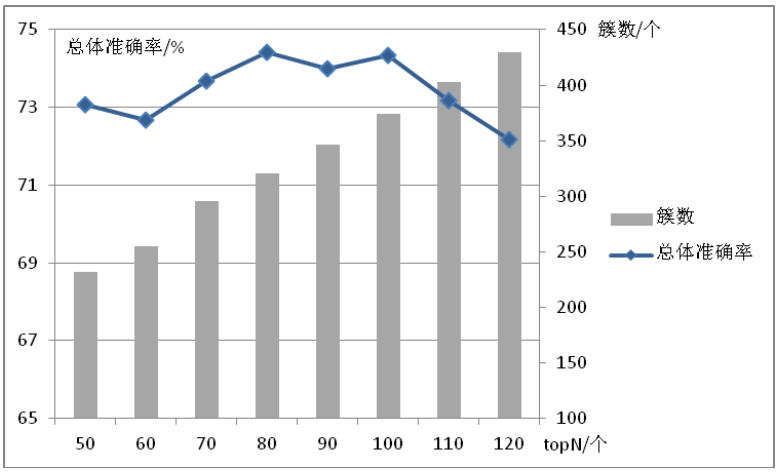

Fig. 4 Experimental results with different topN

图 4 取不同 topN 的实验结果

以 *word2vec* 的参数 *size* 值为 200 时，将 3.3 节提取相似特征过程中的参数 *topN* 取值为 50~120，以 10 为间隔进行实验，得到的实验结果如图 4 所示。随着 *topN* 的值变大，提取到的簇数目逐渐增加。因为簇的数目在种子特征数目不变时，由提取到的相似特征数目决定，所以当 *topN* 取值越大时，与情感特征匹配的相似特征词越多，相应地提取到的簇数目越多。另外，从图 4 可知，随着 *topN* 值的增加，分类准确率有所提高。在 *topN* 取值为 80~100 时，情感分类效果较好，分类准确率相对较稳定。但是当 *topN* 值大于 100 时，分类准确率反而下降。因此，取合适的 *topN* 值对提高分类准确率有一定帮助。

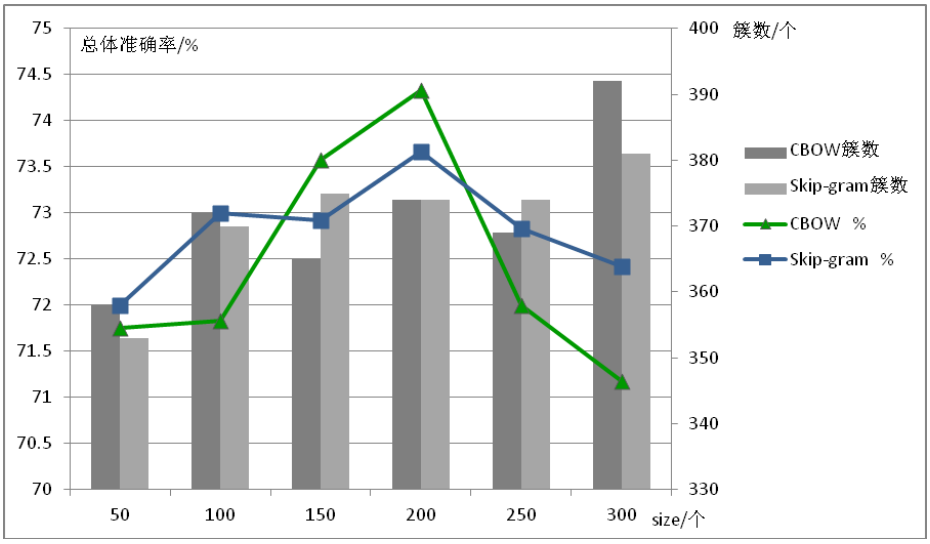

Fig. 5 Experimental results with different parameters of word2vec

图 5 word2vec 取不同参数值的实验结果

将 *topN* 取值 100，*word2vec* 参数中的 *size* 取为 50~300 维，以 50 作为间隔，参数 *isCbow* 分别取 *CBOW* 算法和 *Skip-gram* 算法训练词向量模型，得到的情感分类实验结果如图 5 所示。*CBOW* 簇数和 *Skip-gram* 簇数分别表示在两个不同模型的情况下，提取得到的特征簇的数目，实验过程中词向量的维数从 50 增加到 200 时，提取得到的特征簇的数目跟着变化，并且参数 *isCbow* 不管是选用 *CBOW* 算法还是 *Skip-gram* 算法，总体分类准确率都有所提高，但是随着维数的继续增加，总体分类准确率反而会下降。在维数取到 200 时，*CBOW* 算法对应的分类器取到最高的准确率 74.33%。另外，*Skip-gram* 对应的分类准确率方差为 0.27，低于 *CBOW* 对应的分类准确率方差为 1.25，说明随着维数 *size* 的变化，*Skip-gram* 对应的分类器在微博情感分类上的稳定性优于 *CBOW* 算法对应的分类器。

（2）情感特征降维对比实验

以 *topN* 值为 100，*word2vec* 的 *size* 设为 200，选用 *CBOW* 算法，进行实验 *A*、*B*、*C*，对比分析实验结果，如表 3 所示。实验 *A* 选取所有情感特征构建文本向量。实验 *B* 利用 *fastPCA* 算法提取对实验 *A* 中的情感特征进行降维。实验 *C* 采用本文构造特征簇的方法对情感特征进行降维。特征项的权重值采用 *tfidf* 值，由公式 *tfidf=TF*IDF* 计算得到，*TF* 为代表词频，*IDF* 代表逆向文档频率。

Table 4 Experimental results

表 4 实验结果（%）

实验	特征	特征维数	负向准确率	正向准确率	总体准确率
A	所有情感特征	3186	70.17	66.50	68.33
B	*fastPCA* 降维	374	72.833	72.67	72.75
C	特征簇	374	76.67	72	74.33

由表 4 得到，与实验 *A* 和实验 *B* 对比，实验 *C* 使用本文提出的方法在微博情感分类准确率上都有提高，说明了本文提出方法是有效的。实验 *C* 取得的分类准确率比实验 *A* 高 6%，说明本文基于特征簇构造情感特征的方法比简单使用所有情感特征表示文本向量更有效，而且本文构造的文本向量的维数为 374，即特征簇的数目，相比于实验 *A* 中 3186 的维数降低了 88.27%，使得文本方法在大规模语料的分析任务上更有优势。实验 *B* 利用 *fastPCA* 方法对情感特征进行主成分分析，实现与实验 *C* 同样程度的降维效果，从表 4 的实验结果可知，实验 *C* 相比实验 *B* 的分类总体准确率较高，说明本文方法在特征降维方面表现更出色，降维得到的特征项可以有效应用到微博文本情感分类中。

5 结论和将来的研究工作

本文针对微博短文本特征规模大以及自身特征少的特点，提出一种将情感特征合并成特征簇的方法，实现情感特征的降维，规避数据稀疏性问题。本文方法引入语义学习模型，学习词语之间潜在的语义关联，利用 *word2vec* 训练大规模语料库，得到包含语义信息的词向量模型，把单个词语表示成多维向量的形式。在基于词向量计算词语相似度方法的基础上，结合情感词典，设计算法将情感特征合并成特征簇，以此构造出维数较低的文本向量，并作为机器学习分类方法的输入，构建情感分类器对微博短文本的情感进行分类。实验验证，本文方法对情感特征的降维是可行和有效的，并且取得很好的分类效果。

语料库的规模与内容对 *word2vec* 学习词向量模型有一定的影响，同时微博文本表意多样化，存在大量的网络新词、表情符等特殊符号，都会影响分类的效果。因此文本的下一步工作是扩展微博语料库，学习更丰富的微博文本特征，提高微博情感分类的准确率。

致谢 在此,我们向对本文的工作给予支持和建议的专家表示衷心的感谢。

参考文献:

[1] 肖永磊, 刘盛华, 刘悦, 等. 社会媒体短文本内容的语义概念关联和扩展[J]. 中文信息学报, 2014, 28(4): 21-28.

[2] 梁军,柴玉梅,原慧斌,昝红英,刘铭. 基于深度学习的微博情感分析[J]. 中文信息学报,2014,05:155-161.

[3] 杨佳能,阳爱民,周咏梅. 基于语义分析的中文微博情感分类方法[J]. 山东大学学报(理学版),2014,11:14-21+30..

[4] 贺飞艳,何炎祥,刘楠,刘健博,彭敏. 面向微博短文本的细粒度情感特征抽取方法[J]. 北京大学学报(自然科学版),2014,01:48-54.

[5] 庞磊,李寿山,周国栋. 基于情绪知识的中文微博情感分类方法[J]. 计算机工程,2012,13:156-158+162.

[6] 张珊,于留宝,胡长军. 基于表情图片与情感词的中文微博情感分析[J]. 计算机科学,2012,S3:146-148+176.

[7] 吴方照,王丙坤,黄永峰. 基于文本和社交语境的微博数据情感分类[J]. 清华大学学报(自然科学版),2014,10:1373-1376+1383.

[8] 周剑峰,阳爱民,周咏梅,王璇璇. 基于二元搭配词的微博情感特征选择[J]. 计算机工程,2014,06:162-165

[9] Yamamoto Y, Kumamoto T, Nadamoto A. Role of Emoticons for Multidimensional Sentiment Analysis of Twitter[C]//Proceedings of the 16th International Conference on Information Integration and Web-based Applications & Services. ACM, 2014: 107-115.

[10] Khan F H, Bashir S, Qamar U. TOM: Twitter opinion mining framework using hybrid classification scheme[J]. Decision Support

Systems, 2014, 57: 245-257.

[11] 王磊,苗夺谦,张志飞,余鹰. 基于主题的文本句情感分析[J]. 计算机科学,2014,03:32-35.

[12] 杨震,赖英旭,段立娟,李玉鑑. 基于上下文重构的短文本情感极性判别研究[J]. 自动化学报,2012,01:55-67.

[13] 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.

[14] Mikolov T, Yih W, Zweig G. Linguistic Regularities in Continuous Space Word Representations[C]//HLT-NAACL. 2013: 746-751.

[15] 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.

[16] Sharma A, Paliwal K K. Fast principal component analysis using fixed-point algorithm[J]. Pattern Recognition Letters, 2007, 28(10): 1151-1155.