



基于特征融合的层次结构微博情感分类

朱宪莹,刘箴,金炜,刘婷婷,刘翠娟,柴艳杰

(宁波大学信息科学与工程学院,浙江 宁波 315211)

摘要:情感分类是观点挖掘的热点研究之一,微博文本情感分类具有很高的应用价值。鉴于传统特征选择方法存在语义缺陷,采用神经网络语言模型,提出了基于概率模型的对词向量进行权重分配的深层特征表示方法,构建文本语义向量。将文本深层特征与浅层特征融合,构建融合语义信息的特征向量,弥补传统特征选择方法语义的缺陷。采用SVM层次结构分类模型,实现多种情感分类。实验结果表明,采用特征融合的层次结构情感分类方法,能有效提高微博情感分类的准确率。

关键词:情感分类;词向量;深层特征;特征融合;层次结构分类模型

中图分类号:TP391

文献标识码:A

doi: 10.11959/j.issn.1000-0801.2016182

Hierarchical micro-blog sentiment classification based on feature fusion

ZHU Xianying, LIU Zhen, JIN Wei, LIU Tingting, LIU Cuijuan, CHAI Yanjie

Faculty of Information Science and Technology, Ningbo University, Ningbo 315211, China

Abstract: Sentiment classification is an important issue of opinion mining. It has a high application value to classify sentiment in micro-blogs. As traditional feature selection method has semantic gap, a neural network language model was used to propose a deep feature representation method based on probability model to distribute weight to the word vector. Using this method, text semantic vector could be built. In order to avoid the semantic gap, the deep features and shallow features of text were integrated and feature vector that contained semantic information was constructed. With SVM hierarchical classification model, a variety of sentiments could be classified. Experimental results show that the hierarchical sentiment classification method based on feature fusion can improve the accuracy of sentiment classification in micro-blogs.

Key words: sentiment classification, word vector, deep feature, feature fusion, hierarchical classification model

1 引言

近年来,随着各种网络社会媒体的出现,在众多社会

媒体中挖掘用户情感已成为自然语言处理和数据挖掘领域中的重要内容。微博因其开放性、及时性、广泛性等特点,自问世以来,吸引了众多研究人员的关注。针对微博的

收稿日期:2016-06-16;修回日期:2016-07-05

基金项目:国家自然科学基金资助项目(No.61373068, No.61271399);宁波市科技计划基金资助项目(No.2015A610128, No.2015C50053, No.2015D10011, No.2011B81002);高等学校博士学科点专项科研基金资助项目(No.20133305110004)

Foundation Items: The National Natural Science Foundation of China(No.61373068, No.61271399), Ningbo Science and Technology Plan Project(No.2015A610128, No.2015C50053, No.2015D10011, No.2011B81002), Specialized Research Fund for the Doctoral Program of Higher Education(No.20133305110004)

自然语言处理研究已成为当前的热门,其中,情感分析技术是该领域的重要研究内容。

传统文本情感分类中,研究的主要内容是特征选择方法。传统的特征选择方法选取的特征均为文本浅层特征,没有考虑文本的语义信息。2013年Google(谷歌)推出了Word2Vec,该工具是一款开源框架,其利用深度学习思想,采用神经网络语言模型,通过训练,把对文本内容的处理简化为 K 维向量空间中的向量运算。Word2Vec利用词的上下文,使语义信息更加丰富。本文采用神经网络语言模型,利用Word2Vec工具训练词向量,提出WWBP方法(word vector weighting distribution based on probabilistic model of deep features representation,基于概率模型的词向量权重分配的深层特征表示方法),该方法通过概率模型,建立文本概率向量,采用softmax函数归一化文本概率向量,结合文本词向量矩阵与情感系数,构建文本语义向量;将文本深层特征与浅层特征融合,构建融合语义信息的特征向量,弥补传统特征选择方法的语义缺陷。

传统文本情感分类主要是粗粒度的分类,如极性分类或多种情感分类。但语料中无情感的文本往往占很大比例,会降低分类器的精度。在多分类中,情感之间存在一定的相关性,采用全局分类器难以达到很好的分类效果。基于以上问题,提出基于SVM的层次结构情感分类模型。

本文采用WWBP方法与传统特征选择方法融合的方法,构建含有语义信息的特征向量,采用层次结构的SVM分类器,实现微博情感的多分类。

2 相关工作

基于机器学习的文本情感分析技术在情感分析领域取得了很大的发展。Pang^[1]首次将机器学习的方法应用于文本情感分析中,尝试使用N-gram词语特征和词性特征,并对比了朴素贝叶斯、最大熵、SVM这3种分类器,实现对电影评论的分析,实验结果表明SVM分类器分类效果最好。2004年,Pang^[2]将主、客观句的总结引入情感分类中,通过最小分割法选择文本中的主观句,对选择出的主观句进行训练和测试,对文本进行情感分类,提高了分类器的准确度。Alam等人^[3]对神话故事进行情感分析,使用情感词特征、POS特征、特殊标点符号特征,根据神话故事自身特点,抽取特定的文本特征,利用SVM分类器对文本进行分类,实现6种基本情绪分类^[4]。Ghazi等人^[5]采用两层的层

次模型,克服了数据集不平衡的状况,结果表明层次结构分类器模型提高了分类的准确率。之后他们又提出了两层和三层的层次结构分类模型^[6]。

Huang等人^[7]提出一种多标签多任务的情感分类模型,应用在情感分类和话题分类中,解决了文本多情感问题。Liu等人^[8]提出基于多标签的情感分析,首次将多标签的分类方法应用到微博情感分析中。Xu等人^[9]提出一种基于层次情感分类的中文微博情感分析模型,实现细粒度的微博情感分类,但只考虑了浅层的文本特征,且情感类型存在冗余。Cho等人^[10]构建了一个领域情感词典,通过情感词典构建特征向量,将情感词作为特征,采用SVM分类器进行情感分类;实验证明,该方法提高了分类的准确率。Sun等人^[11]利用微博中的表情符号,采用SVM分类器进行情感分类,实现微博情感分类。刘翠娟等人^[12]采用基于依存句法和人工标注相结合的情感分析技术,对新浪微博的话题数据进行群体情感强度分析,并对微博情感分析进行了可视化。

Hinton^[13]在1986年提出概念的分布式表达,开创了词语分布式表达的先河。其基本思想是通过大量语料库训练,将某种语言中的每个词语映射成一个固定长度的向量,得到词向量表示方法。Word2Vec^[14,15]由Google团队提出并实现,该工具能够在较短时间内,从大规模语料库中学习到高质量的词向量。

本文提出基于概率模型的词向量权重分配的深层特征表示方法(WWBP方法),将文本浅层特征与深层特征融合,采用层次结构分类器,实现微博情感分类。

3 情感分类方法设计

本节从传统文本特征、基于词向量的文本深层特征、特征融合和分类器构建几个方面进行阐述。基于特征融合的层次结构情感分类总体框架如图1所示,主要有数据采集、数据预处理、特征选择、深层特征表示、层次结构分类器构建这几部分。其中,深层特征表示和层次结构分类器的构建为主要创新部分。

3.1 传统文本特征

传统的文本情感分析特征主要有以下几种。

(1) 词典特征

包括情感词特征、词特征或短语特征。本文利用大连理工大学的情感词典构建情感词特征,该情感词典囊括7类情感,即乐、好、怒、哀、惧、恶、惊。

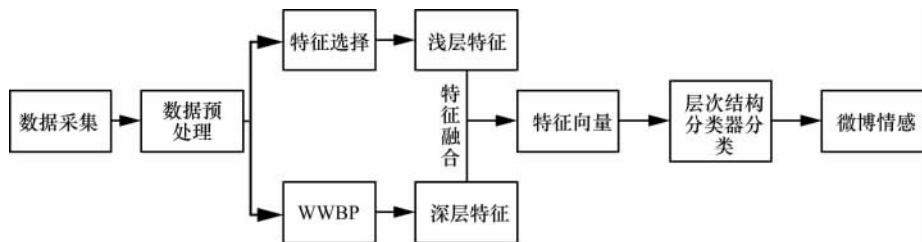


图1 基于特征融合的层次结构情感分类框架

(2) 语言结构特征

包括 N-gram 特征、词性(POS)特征。N-gram 特征有 unigrams、bigrams、trigrams 特征, 本文选取了 unigrams 和 bigrams 特征, 词性特征选取了名词、动词、形容词等特征。

(3) 微博特征

包括特殊标点符号特征、表情符号特征。特殊标点符号, 如连续出现的“!”“?”, 该特征在情感表达中起强调作用。表情符号在微博中直接表达用户情感, 是评判微博情感的重要指标。

本文利用向量空间模型与布尔权重表示法构建基于传统特征选择方法的文本特征向量。分别使用 CHI(Chi-square, 卡方统计)、IG (information gain, 信息增益)、MI (mutual information, 互信息) 这几种特征选择方法, 实现特征选择。

3.2 基于词向量的文本深层特征

在自然语言处理领域, 深度学习的一个重要应用是挖掘文本特征的分布式表示, 深度学习将单词用固定长度的实数向量表示, 即词向量, 词向量包含了该词在文本中的语义信息。传统特征选择方法存在语义缺陷, 只考虑文本的浅层特征, 没有表达词之间、句子之间深层的语义信息。本文从深度学习文本表示的角度, 针对传统文本特征存在语义缺陷的问题, 提出 WWBP 方法。

定义 1 (全局概率模型) 指每个词或短语在语料库中的贡献度, 频率高的贡献度大, 反之贡献度小。采用词或短语在语料中的概率作为贡献度的评价标准, 提出了全局概率模型, 如下:

$$global(p) = \left[\frac{sum_{w_1}}{sum_{all}} \quad \frac{sum_{w_2}}{sum_{all}} \quad \dots \quad \frac{sum_{w_n}}{sum_{all}} \right] \quad (1)$$

其中, sum_{w_i} 表示词 w_i 在语料中出现的次数, sum_{all} 表示语料中词的总数, sum_{w_i}/sum_{all} 表示词 w_i 在语料中出现的概率, 即该词在语料库中的全局概率。该模型描述了词 w_i 在语料中的贡献程度。

本文提出的 WWBP 方法利用全局概率模型, 结合词向量, 构建与词向量维度相同的文本语义向量, 其实现流

程如下。

(1) 数据预处理

使用新浪微博 API 采集微博数据, 对采集的数据进行预处理, 如去除特殊符号、网址、话题符号等。利用中国科学院分词系统, 对微博数据分词, 得到 2 GB 微博数据语料库。将该语料库作为训练词向量的语料库。

(2) 词向量训练

利用 gensim 库的 Word2Vec 工具, 训练微博数据语料库的词向量, 得到每个词的词向量。本文分别训练了 100、150、200、250 和 300 维的词向量。

(3) 概率语料库构建

通过全局概率模型式(1), 计算语料库中词的全局概率, 构建概率语料库。该概率语料库共 98 724 个词, 每一行由词和其对应的概率组成。

(4) 文本概率向量

结合概率语料库与词向量, 构建文本概率向量。文本概率向量表示文本中词的贡献度分布, 其计算式为:

$$P_j = [p_1 \ p_2 \ \dots \ p_n] \quad (2)$$

其中, P_j 表示第 j 篇文本的概率向量, $p_i (i=1, 2, \dots, n)$ 表示第 j 篇文本中第 i 个词的概率, n 表示第 j 篇文本词的数目。

在概率语料库中词的概率成指数形式, 对语料贡献度不明显。使用该概率为词向量分配权重, 使得词向量每个维度语义的信息较少, 构建的文本语义向量语义信息不明显。为了使文本语义向量包含更多的语义信息, 采用基于 softmax 函数的概率向量归一化方法, 利用 softmax 函数柔化文本概率向量, 将文本概率向量每个维度的概率值归一化为 0~1 之间的数值, 使出现在文本中词的概率和为 1。其计算式为:

$$\sigma(p_i) = \frac{e^{p_i}}{\sum_{k=1}^n e^{p_k}} \quad (3)$$

其中, $p_i (i=1, 2, \dots, n)$ 表示该篇文本中第 i 个词的概率, n 表示文本中词的总数。通过 softmax 函数归一化, 得到新

的文本概率向量为:

$$P_j' = [\sigma(p_1) \sigma(p_2) \cdots \sigma(p_n)] \quad (4)$$

其中, P_j' 表示第 j 篇文本归一化的概率向量, $\sigma(p_i)$ ($i=1,2,\cdots,n$) 表示第 j 篇文本中第 i 个词归一化后的概率,

且 $\sum_{i=1}^n \sigma(p_i)=1$ 。文本概率向量归一化后, 新的文本概率向量数
值间的差距缩小, 更能反映每个词对文本的贡献度。

(5) 文本词向量矩阵

一条微博由多个词构成, 每个词均有唯一的词向量与
之对应。文本词向量式为:

$$M_j = \begin{bmatrix} v_{11} & \cdots & v_{1k} \\ \vdots & & \vdots \\ v_{n1} & \cdots & v_{nk} \end{bmatrix} = [\vec{w}_1 \quad \vec{w}_2 \quad \cdots \quad \vec{w}_n]^T \quad (5)$$

其中, M_j 表示第 j 篇文本的词向量矩阵, v_{ik} ($i=1,2,\cdots, n, k=1,2,\cdots,K$) 表示第 i 个词向量第 k 维的语义信息, $\vec{w}_i$ 表示第 j 篇文本中第 i 个词的词向量。根据式(4)得到的文本
概率向量, 为文本词向量分配权值。权值分配后的文本词
向量矩阵为:

$$M_j' = [\sigma(p_1) \cdot \vec{w}_1 \quad \sigma(p_2) \cdot \vec{w}_2 \quad \cdots \quad \sigma(p_n) \cdot \vec{w}_n]^T \quad (6)$$

其中, $\sigma(p_i)$ ($i=1,2,\cdots,n$) 表示词向量 $\vec{w}_i$ 对应的归一化后
的概率。

(6) 情感系数设定

微博中的情感词直接表达微博情感, 为了强调情感词
对微博情感的影响, 设定情感系数 α , 其中, $\alpha \in (0,1)$ 。通常
非情感词在微博中对情感的影响比情感词小, 为了强调非
情感词与情感词对文本情感贡献的不同, 设定非情感系数
为 β , 其中, $\beta \in (0,1)$ 。对微博中每个词设定情感系数, 设定
规则为:

$$\begin{cases} \alpha \cdot \vec{w}_i, & w_i \in \text{emotion} \\ \beta \cdot \vec{w}_i, & w_i \notin \text{emotion} \end{cases} \quad (7)$$

其中, $\vec{w}_i$ 表示词 w_i 的词向量, emotion 表示情感词典。
令 $\alpha+\beta=1$, 初始化情感系数 $\alpha=0, \beta=1.0$, α 以步长 0.1 的速
度递增, 由 $\alpha+\beta=1$, 得到非情感系数 β ; 根据情感分类结
果不断调整 α, β , 直至得到最佳分类效果。采用上述情感
系数分配规则, 突出了情感词与非情感词对微博情感贡献
度的不同。

(7) 文本语义向量

将文本词向量矩阵中所有词向量对应维度相加, 构建
的文本语义向量为:

$$D_j = \left[\sum_{i=1}^n \sigma(p_i) \cdot v_{i1} \quad \sum_{i=1}^n \sigma(p_i) \cdot v_{i2} \cdots \sum_{i=1}^n \sigma(p_i) \cdot v_{iK} \right] \quad (8)$$

其中, $\sum_{i=1}^n \sigma(p_i) \cdot v_{ij}$ ($j=1,2,\cdots,K$) 表示文本词向量矩阵 M_j

对应列的和, 该向量表征了文本的语义信息。

3.3 特征融合

特征选择方法有 CHI、IG、MI 等, 每种特征选择方法
都有其自身的优缺点。CHI 判定了特征项与类别之间的关
联程度, 关联度越大, 信息量也越多; 但该方法只统计文本
中是否出现词, 而不管词出现的次数, 因此对低频词有所
偏袒。IG 是基于信息熵的特征选择方法, 文本特征越是有
序, 信息熵就越低; 反之, 一个系统越是混乱, 信息熵越高,
信息熵可以更好地区分特征对于类别区分度的贡献。但是
该方法只考察特征对整个系统的贡献, 而不能具体到某个
类别中。MI 作为特征词和类别之间的测度, 互信息量越
大, 其特征类别相关度越高, 该方法不需要对特征词和类
别之间关系的性质作任何假设, 适合于文本分类的特征选
择; 但是该方法存在词频缺陷。

由以上可知, CHI、IG、MI 方法仅选取与类别相关的特
征项, 没有考虑文本词之间、上下文之间的语义关系, 存在
语义缺陷问题。WWBP 方法结合概率模型与词向量, 抽取
文本的语义信息, 构建文本语义向量; CHI、IG、MI 方法选
取与类别相关的特征项, 构建文本特征向量; 因此本文提
出将语义向量与特征向量融合的方法, 构建融合语义信息
的文本特征向量。采用的特征融合方式分别为 CHI+
WWBP、IG+WWBP、MI+WWBP。

3.4 层次结构分类模型

传统分类模型实现的是粗粒度的情感分类, 没有考虑
文本之间的情感联系。例如: a. 今天很开心; b. 今天很伤心;
c. 今天天气晴朗。a 是积极的文本, 是开心的情感; b 是消
极的文本, 是伤心的情感; c 是没有情感的文本。文本数据
量大时, 没有情感的文本会降低传统分类器的精度, 且有
情感的文本, 由于极性不同, 也会降低情感分类的精度。
因此本文构建了一个基于 SVM 的层次结构分类器模型,
首次使用层次结构分类模型实现 7 种情感分类, 提高了情
感分类的精度。

本文构建了具有 3 个层次结构的情感分类模型。第一
层构建主客观分类器, 实现主客观分类, 剔除文本中无情
感的文本, 减少最终情感分类的误判率。第二层构建极性
分类器, 将第一层分类得到的有情感的文本作为第二层的



输入数据,实现文本的极性分类。第三层分别构建消极情感分类器和积极情感分类器,将第二层分类后极性不同的文本分别作为第三层的输入数据,得到文本最终的情感类型。层次结构情感分类模型如图2所示。

4 实验语料与评价指标

4.1 实验语料

本文利用自然语言处理与中文计算会议数据集,将近几年该会议的数据集进行整合,得到14 000条已经标注情感类型的微博数据。主要的情感类型有无情感、厌恶、伤心、生气、吃惊、害怕、喜好、开心。各情感类型分布见表1。

表1 各情感类型微博数量

情感类型	微博数目/条
无情感	7 490
开心	1 310
喜好	1 879
厌恶	1 239
伤心	1 018
生气	595
吃惊	324
害怕	145

对该数据集(data)进行情感层次划分,首先,划分为有情感类型的文本数据(subject)和无情感类型的文本数据(object)。其次,对有情感数据集进行极性划分,将喜好和开心划分为积极(positive)情感,其他情感类型,包括厌恶、伤心、生气、吃惊、害怕划分为消极(negative)情感。

将数据集划分为10 000条数据的训练集和4 000条数据的测试集,用来训练全局分类器;将data、subject、positive和negative数据集分别划分为训练集和测试集,用来训练层次结构分类器。

4.2 评价指标

本文采用 P (precision,准确率)、 R (recall,召回率)、 $F1$

($F1$ -measure, $F1$ 值)作为评价指标。

文本类别归属判别见表2。

表2 文本类别归属判别

文本类别	实际属于 c_i 类的文本数	实际不属于 c_i 类的文本数
分类器判别为 c_i 的文本数	A	B
分类器判别不为 c_i 的文本数	C	D

(1) 准确率

分类器判别为 c_i 类别的文本数与实际属于 c_i 类别的文本数的比值,其计算式为:

$$P = \frac{A}{A+B} \quad (9)$$

(2) 召回率

实际属于 c_i 类别的文本数与分类器判别为 c_i 类别的文本数的比值,其计算式为:

$$R = \frac{A}{A+C} \quad (10)$$

准确率和召回率分别反映分类器不同方面的性能:准确率反映分类器的准确性,召回率反映分类器的完备性。

(3) $F1$ 值

它综合了准确率和召回率,其计算式为:

$$F1 = \frac{2 \times P \times R}{P + R} \quad (11)$$

5 实验结果及分析

实验采用 LibSVM 工具包和已处理好的实验语料。

5.1 基于 WWBP 方法的实验

为了验证 WWBP 方法的有效性,做了以下实验。首先,研究了不同维度的词向量构建的文本语义向量,对有无情感分类、极性分类和多分类实验结果的影响,选择最佳的维度进行 WWBP 方法与传统方法 CHI、IG、MI 的对比

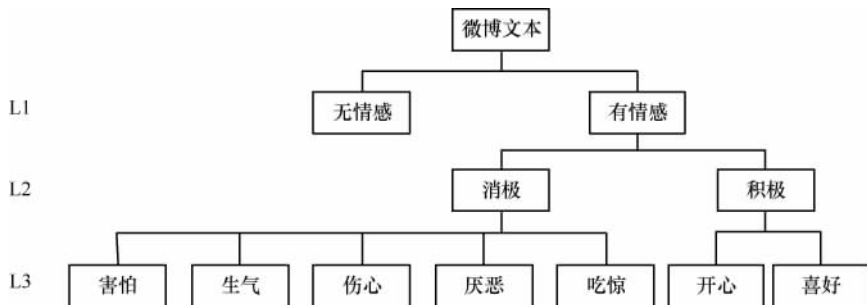


图2 层次结构情感分类模型

实验,由于 MI 方法存在词频缺陷,根据词频大小对选取的特征项进行排序,提取词频较大的特征项,计算它的互信息。其次,为了得到更好的分类效果,提出了特征融合的方法,将 CHI、IG、MI 方法分别与 WWBP 方法融合,构建包含语义信息的特征向量,即 CHI+WWBP、IG+WWBP、MI+WWBP,分别实现文本有无情感分类、极性分类、多种情感分类。

用 2 GB 的微博语料训练 50、100、150、200、250、300 维词向量,分别用这几种维度的词向量,根据深层文本特征构建文本语义向量,实现有无情感分类(motion)、极性分类(polarity)、多种情感分类(positive 和 negative),实验结果如图 3 所示。

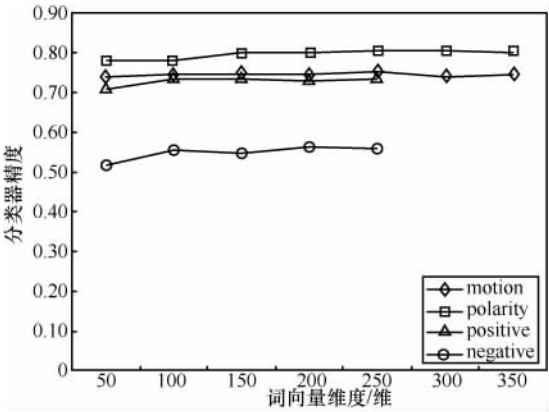


图 3 不同维度词向量分类器精度

实验结果表明,随着词向量维度的增加,分类器精度不是线性变化,当词向量的维度由 50 维增加为 100 维时,每个层次的分类精度均有所提高;维度继续提升时,精度没有明显的变化。因此,在接下来的实验中,采用 100 维的词向量。

5.1.1 有无情感分类

数据集采用 data,分类器采用 SVM,特征选择方法采

用 CHI、IG、MI、WWBP 和特征融合方法。实验结果见表 3。

实验结果表明,在有无情感的二分类中,WWBP 方法的 P 、 R 、 $F1$ 值均比传统文本特征选择方法高 1%~5%。特征融合方法与传统方法相比,其 P 、 R 、 $F1$ 值平均提高了 3%左右。WWBP 方法和特征融合方法与传统方法相比,分类精度均有提高,如图 4 所示。

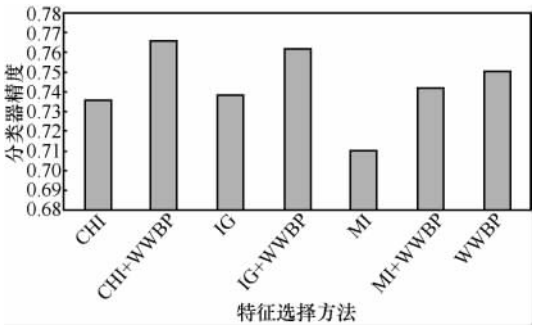


图 4 基于各特征选择方法的有无情感分类器精度

5.1.2 极性分类

数据集采用 subject,分类器采用 SVM,特征选择方法采用 CHI、IG、MI、WWBP 和特征融合方法。实验结果见表 4。

实验结果表明,在极性分类中,WWBP 方法的 P 、 R 、 $F1$ 值均高于 MI 方法;积极文本分类时,WWBP 方法的 R 值高于其他特征选择方法;消极文本分类时,WWBP 方法的 P 值高于其他传统特征选择方法, $F1$ 值高于 MI 和 CHI 方法。特征融合方法与传统方法相比,其 P 、 R 、 $F1$ 值平均提高了 1%左右。WWBP 方法得到的分类器精度高于 IG 和 MI 方法,特征融合的方法与传统方法相比,分类器精度均得到提高,具体如图 5 所示。

5.1.3 多种情感分类

数据集采用 negative,分类器采用 SVM,特征选择方法采用 CHI、IG、MI、WWBP 和特征融合方法。实验结果见表 5。

表 3 基于各特征选择方法的有无情感分类

特征选择方法	有情感			无情感		
	P	R	$F1$	P	R	$F1$
CHI	0.746	0.759	0.753	0.709	0.716	0.716
CHI+WWBP	0.777	0.783	0.780	0.754	0.747	0.750
IG	0.744	0.772	0.758	0.732	0.701	0.716
IG+WWBP	0.770	0.785	0.777	0.753	0.736	0.744
MI	0.744	0.754	0.734	0.704	0.661	0.682
MI+WWBP	0.752	0.765	0.759	0.731	0.716	0.723
WWBP	0.762	0.767	0.765	0.750	0.736	0.733



表 4 基于各特征选择方法的极性分类

特征选择方法	积极			消极		
	<i>P</i>	<i>R</i>	<i>F1</i>	<i>P</i>	<i>R</i>	<i>F1</i>
CHI	0.822	0.752	0.786	0.794	0.805	0.799
CHI+WWBP	0.822	0.794	0.808	0.812	0.838	0.825
IG	0.847	0.785	0.815	0.810	0.866	0.837
IG+WWBP	0.848	0.794	0.820	0.817	0.866	0.841
MI	0.806	0.751	0.778	0.780	0.830	0.804
MI+WWBP	0.816	0.776	0.795	0.798	0.834	0.816
WWBP	0.809	0.801	0.805	0.814	0.822	0.818

表 5 基于各特征选择方法的消极情感多分类

特征选择方法	厌恶			伤心			生气			害怕		
	<i>P</i>	<i>R</i>	<i>F1</i>	<i>P</i>	<i>R</i>	<i>F1</i>	<i>P</i>	<i>R</i>	<i>F1</i>	<i>P</i>	<i>R</i>	<i>F1</i>
CHI	0.551	0.771	0.643	0.590	0.722	0.649	0.632	0.326	0.430	0.533	0.133	0.213
CHI+WWBP	0.564	0.739	0.640	0.606	0.765	0.676	0.612	0.353	0.447	0.530	0.150	0.234
IG	0.516	0.788	0.623	0.580	0.742	0.651	0.663	0.236	0.349	1.000	0.033	0.065
IG+WWBP	0.540	0.773	0.636	0.590	0.780	0.676	0.655	0.295	0.406	0.857	0.100	0.179
MI	0.448	0.877	0.593	0.631	0.537	0.580	0.693	0.202	0.312	1.000	0.017	0.033
MI+WWBP	0.515	0.829	0.636	0.623	0.752	0.681	0.703	0.248	0.367	1.000	0.033	0.065
WWBP	0.532	0.752	0.623	0.625	0.750	0.681	0.554	0.260	0.354	0.000	0.000	0.000

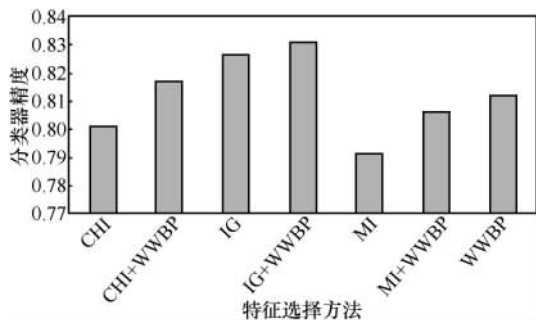


图 5 基于各特征选择方法的极性分类器精度

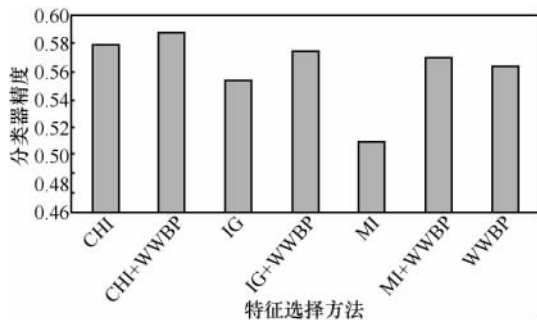


图 6 基于各特征选择方法的消极情感分类器精度

实验结果表明,在多情感分类中,WWBP方法的 *F1* 值高于 IG、MI 方法;伤心分类时,WWBP 方法分类效果最优,其 *P*、*R*、*F1* 值均高于传统方法。特征融合方法与传统方法相比,其 *P*、*R*、*F1* 值均有所提高。WWBP 方法得到的分类器精度高于 IG 和 MI 方法,特征融合法与传统方法相比,分类器精度均得到提高,具体如图 6 所示。

以上实验得出,WWBP 方法的分类效果优于传统的特征选择,且将 WWBP 方法与传统方法融合,在有无情感分类、极性分类和多分类中分类效果均得到一定提高,取得了更好的分类效果。

5.2 基于特征融合的层次结构情感分类实验

由上述实验结果可知,WWBP 方法与传统方法融合实

现的分类效果优于传统方法。为了验证基于特征融合的层次结构情感分类方法的高效性,分类器采用全局 SVM 分类器和基于 SVM 的层次结构分类器,特征选择方法采用 CHI、IG、MI、WWBP 和特征融合方法,对微博进行多种情感分类。实验结果见表 6、表 7。

由表 6、表 7 可知,采用 SVM 的层次结构分类器,对微博进行 8 种情感分类,与全局分类器相比,其 *P*、*R*、*F1* 值均得到了一定幅度的提高。对无情感的微博分类,其 *P*、*R*、*F1* 值达到 80% 以上。并且每种特征选择方法,使用层次结构分类器后,其 *P*、*R*、*F1* 值均得到了提高。实验结果表明,采用层次结构分类器模型时,根据每个层次分类任务使用对应分类器, *P*、*R*、*F1* 值均得到了提高。

表 6 基于各特征选择方法的全局情感多分类

情感类型	CHI			CHI+WWBP			IG			IG+WWBP			MI			MI+WWBP		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
中性	0.641	0.905	0.751	0.675	0.906	0.774	0.610	0.948	0.742	0.640	0.922	0.755	0.590	0.955	0.729	0.630	0.935	0.752
开心	0.462	0.336	0.389	0.50	0.419	0.456	0.498	0.289	0.366	0.49	0.266	0.345	0.50	0.198	0.283	0.460	0.216	0.295
喜好	0.447	0.272	0.337	0.50	0.313	0.379	0.507	0.214	0.301	0.496	0.389	0.436	0.457	0.160	0.236	0.523	0.372	0.435
厌恶	0.477	0.209	0.290	0.429	0.214	0.286	0.420	0.097	0.158	0.417	0.157	0.228	0.386	0.071	0.121	0.431	0.143	0.215
伤心	0.462	0.154	0.231	0.551	0.270	0.362	0.420	0.066	0.114	0.517	0.144	0.226	0.409	0.028	0.053	0.513	0.125	0.202
生气	0.488	0.275	0.352	0.495	0.315	0.385	0.640	0.215	0.322	0.557	0.262	0.356	0.611	0.222	0.325	0.535	0.255	0.346
吃惊	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
害怕	0.667	0.057	0.105	1.000	0.057	0.108	0.50	0.029	0.054	0.750	0.086	0.154	1.00	0.029	0.056	1.000	0.029	0.056

表 7 基于各特征选择方法的层次结构情感多分类

情感类型	CHI			CHI+WWBP			IG			IG+WWBP			MI			MI+WWBP		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
中性	0.828	0.852	0.84	0.892	0.892	0.892	0.799	0.838	0.818	0.879	0.888	0.883	0.804	0.830	0.817	0.833	0.84	0.836
开心	0.541	0.643	0.747	0.865	0.720	0.786	0.864	0.673	0.756	0.875	0.746	0.806	0.836	0.587	0.690	0.842	0.693	0.761
喜好	0.947	0.629	0.731	0.883	0.668	0.761	0.916	0.572	0.704	0.914	0.619	0.738	0.848	0.563	0.677	0.865	0.614	0.718
厌恶	0.771	0.740	0.755	0.805	0.709	0.754	0.707	0.629	0.666	0.754	0.734	0.744	0.593	0.591	0.592	0.692	0.629	0.659
伤心	0.833	0.627	0.716	0.823	0.655	0.73	0.782	0.461	0.580	0.794	0.605	0.687	0.782	0.370	0.502	0.755	0.599	0.668
生气	0.928	0.604	0.732	0.823	0.655	0.730	0.884	0.410	0.560	0.910	0.544	0.681	0.898	0.356	0.510	0.864	0.470	0.609
吃惊	0.973	0.396	0.563	0.925	0.407	0.565	1.000	0.220	0.360	1.00	0.275	0.430	1.000	0.110	0.198	1.000	0.198	0.330
害怕	0.941	0.457	0.615	0.944	0.486	0.642	1.000	0.229	0.372	0.929	0.372	0.531	1.000	0.229	0.372	1.000	0.257	0.409

采用特征融合的层次结构分类器分类,与使用传统方法的层次结构分类器相比,其分类结果的 P 、 R 、 $F1$ 值均得到一定的提高,与使用传统方法的全局分类器相比,其 P 、 R 、 $F1$ 值均提高了 8%~50%。综合评价各种组合,采用 CHI+WWBP 特征融合的层次结构分类器分类效果最佳。

由以上实验可知,采用层次结构的分类效果优于全局分类,说明对情感类型采用对应类别的分类器分类,能够提高分类的准确率;采用特征融合方法的分类效果优于单个特征选择方法分类效果;采用特征融合的层次结构分类器对文本分类,其分类效果最佳。

6 结束语

文本情感分类在自然语言处理、数据挖掘、社交媒体中的应用越来越广,传统的文本特征选择方法存在语义缺陷,选取的特征没有语义信息。Word2Vec 利用词的上下文关系,训练出的词向量包含丰富的语义信息。因此,本文提出 WWBP 方法,构建文本语义向量,该语义向量为文本深层特征;将深层特征与浅层特征融合,构建融合语义信息的特征向量,弥补传统方法语义缺陷问题。实验结果表明,

特征融合的方法与传统特征选择方法相比,在有无情感分类中,准确率、召回率和 $F1$ 值平均提高了 3%左右;在极性分类中,准确率、召回率和 $F1$ 值平均提高了 1%左右;在多分类中,准确率、召回率和 $F1$ 值均有所提高。在多种情感分类时,提出基于特征融合的层次结构情感分类方法,采用 SVM 层次结构情感分类模型,实验结果表明,该方法与全局分类器相比,其准确率、召回率和 $F1$ 值均提高了 8%~50%,提高了微博多种情感分类的准确率。

本文采用的分类器均为 SVM 分类器,未分析不同分类器分类结果的优劣;词向量的好坏取决于语料库的大小。以后将从分类器选取和语料库扩建两方面着手,优化实验。

参考文献:

[1] PANG B, LEE L, VAITHYANATHAN S. Thumbs up sentiment classification using machine learning techniques [J]. Computer Science, 2009(10): 79-86.
[2] PANG B, LEE L. Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales [J]. Arxiv Cornell University Library, 2004: 115-124, arXiv: cs/0506075v1.



- [3] ALAM C O, ROTH D, SPROAT R. Emotions from text: machine learning for text-based emotion prediction [C]// Conference on Human Language Technology and Empirical Methods in Natural Language Processing, October 6-8, 2005, Vancouver, British Columbia, Canada. New York: ACM Press, 2005: 579-586.
- [4] ECKMAN P. Universal and cultural differences in facial expression of emotion [EB/OL]. [2016-05-02]. https://www.researchgate.net/publication/248224937_Universal_and_cultural_differences_in_facial_expression_of_emotion.
- [5] GHAZI D, INKPEN D, SZPAKOWICZ S. Hierarchical versus flat classification of emotions in text [C]//NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, June 5, 2010, Los Angeles, USA. New York: ACM Press, 2010: 140-146.
- [6] GHAZI D, INKPEN D, SZPAKOWICZ S. Hierarchical approach to emotion recognition and classification in texts [M]. Berlin: Springer, 2010: 40-50.
- [7] HUANG S, PENG W, LI J, et al. Sentiment and topic analysis on social media: a multi-task multi-label classification approach [C]//The 5th Annual ACM Web Science Conference, May 2-4, 2013, Paris, France. New York: ACM Press, 2013: 172-181.
- [8] LIU S M, CHEN J H. A multi-label classification based approach for sentiment classification [J]. Expert Systems with Applications, 2015, 42(3): 1083-1093.
- [9] XU H, YANG W, WANG J. Hierarchical emotion classification and emotion component analysis on chinese micro-blog posts[J]. Expert Systems with Applications, 2015, 42(22): 8745-8752.
- [10] CHO S H, KANG H B. Text sentiment classification for SNS-based marketing using domain sentiment dictionary [C]// 2012 IEEE International Conference on Consumer Electronics (ICCE), Jan 13-16, 2012, Las Vegas, NV, USA. New Jersey: IEEE Press, 2012: 717-718.
- [11] SUN X, LI C, YE J. Chinese microblogging emotion classification based on support vector machine [C]//2014 International Conference on Computing, Communication and Networking Technologies (ICCCNT), July 11-13, 2014, Hefei, China. New Jersey: IEEE Press, 2014: 1-5.
- [12] 刘翠娟, 刘箴, 柴艳杰, 等. 基于微博文本数据分析的社会群体情感可视计算方法研究[J]. 北京大学学报(自然科学版), 2016, 52(1): 178-186.
LIU C J, LIU Z, CHAI Y J, et al. Visual study on calculation method of social groups emotional based on the micro-blog post analysis [J]. Journal of Peking University(Natural Science Edition), 2016, 52(1): 178-186.
- [13] HINTON G E. Learning distributed representations of concepts

[EB/OL]. [2002-08-01]. https://www.researchgate.net/publication/2883217_Learning_Distributed_Representations_of_Concepts

- [14] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space [J]. Computer Science, 2013(9), arXiv: 1301.3781v3.
- [15] MIKOLOV T, SUTSKEVER I, Chen K, et al. Distributed representations of words and phrases and their compositionality[J]. Advances in Neural Information Processing Systems, 2013(26): 3111-3119.

[作者简介]



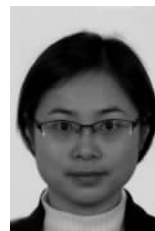
朱宪莹(1991-),女,宁波大学信息科学与工程学院硕士生,主要研究方向为文本情感分析。



刘箴(1965-),男,博士,宁波大学信息科学与工程学院教授,主要研究方向为虚拟现实和社会媒体。



金炜(1969-),男,博士,宁波大学信息科学与工程学院副教授,主要研究方向为图像处理。



刘婷婷(1980-),女,宁波大学信息科学与工程学院博士生,主要研究方向为虚拟现实和社会媒体。

刘翠娟(1979-),女,宁波大学信息科学与工程学院博士生,主要研究方向为社会媒体。

柴艳杰(1968-),女,宁波大学讲师,主要研究方向为信息检索和动漫仿真。