DOI:10.11992/tis.201706026
网络出版地址:http://kns.cnki.net/kcms/detail/23.1538.TP.20171021.1350.010.html

基于语义特征的多视图情感分类方法

吴钟强[1,2],张耀文[1,2],商琳[1,2]
(1.南京大学 计算机软件新技术国家重点实验室,江苏 南京 210046; 2. 南京大学 计算机科学与技术系,江苏 南京 210046)

摘 要:情感分析也称为意见挖掘,是对文本中所包含的情感倾向进行分析的技术。目前很多情感分析工作都是基于纯文本的。而在微博上,除了文本,大量的图片信息也蕴含了丰富的情感信息。本文提出了一种基于文本和图像的多模态分类算法,通过使用潜在语义分析,将文本特征和图像特征分别映射到同维度下的语义空间,得到各自的语义特征,并用 SVM-2K 进行分类。利用新浪微博热门微博栏目下爬取的文字和配图的微博数据进行了实验。实验结果表明,通过融合文本和图像的语义特征,情感分类的效果好于单独使用文本特征或者图像特征。

关键词:情感分析;文本挖掘;潜在语义分析;多模态;语义特征;特征融合;特征提取
中图分类号:TP181 **文献标志码**:A **文章编号**:1673-4785(2017)05-0745-07

中文引用格式:吴钟强,张耀文,商琳. 基于语义特征的多视图情感分类方法[J]. 智能系统学报, 2017, 12(5): 745-751.
英文引用格式: WU Zhongqiang, ZHANG Yaowen, SHANG Lin. Multi-view sentiment classification of microblogs based on semantic features[J]. CAAI transactions on intelligent systems, 2017, 12(5): 745-751.

Multi-view sentiment classification of microblogs based on semantic features

WU Zhongqiang[1, 2], ZHANG Yaowen[1,2], SHANG Lin[1,2]
(1. State Key Laboratory of Novel Software Technology, Nanjing University, Nanjing 210046, China; 2. Department of Computer Science and Technology, Nanjing University, Nanjing 210046, China)

Abstract: The objective in sentiment analysis is to analyze the sentiment tendency contained in subjective text. Most sentiment analysis methods deal with text only and ignore the information provided in the corresponding pictures. In this paper, we propose a multi-view microblog analysis method based on semantic features. Using latent semantic analysis, we map both the text and image features to the semantic space in the same dimensionality, and use SVM-2K to obtain and classify the respective semantic features. We conducted experiments by crawling text and pictures from popular microblogs. The results show that, by combining the semantic features of text and pictures, the sentiment classification result is better than that obtained using text or image features alone.

Keywords: sentiment analysis; text mining; latent semantic analysis; multi-view; semantic features; feature fusion; feature extraction

随着互联网的快速发展,微博自 2006 年以来已经成为社交网络的最主要应用之一。用户可以通过手机或平板电脑等终端设备在微博上发布动态。近年来,从微博数据中挖掘出有价值的信息引起了很多研究者的关注。情感分析或意见挖掘,是一种对人们发表的观点、表达的情感或商品评价进行分析的技术[1]。随着 Pang 等[2]将机器学习方法成功应用在情感分类之后,情感分析领域不断涌现新的工作,针对于粒度的不同可以分为文档级别[2](document level)、句子级别[3](sentence level)和方面级别[4-5](aspect level)。情感分析的应用也越来越广泛,如 Liu 等[6]将其用于预测销售业绩上,Mishne 等[7]使用博文的情感来预测电影票房,O'Connor将文本中的情感与票选关联[8]。但是绝大多数研究都只是基于文本,结合微博图像进行情

收稿日期:2017-06-08. 网络出版日期:2017-10-21.
基金项目:国家自然科学基金项目(61672276);江苏省自然科学基金项目(20161406).
通信作者:吴钟强.E-mail: wuzqchom@163.com.

感分类的工作较少。但图像也是传达情感信息的重要渠道。对于文本和图像并存的情况,图像也可以作为传播情感的载体,如果仅对文本部分进行特征的提取,可能导致对微博整体情感特征提取的缺失,使得整体情感分类的结果不理想。

要使用不同视图的特征就涉及特征融合问题。特征融合被广泛应用在多个领域,如目标跟踪和识别[9]、图像处理[10]等领域,主要可以分为串行融合和并行融合[11]。

本文通过复数矩阵融合的方式并使用潜在语义分析[2](latent semantic analysis, LSA)技术提出了基于语义特征的多视图分类方法。首先,将文本和图像并行融合之后的特征,通过潜在语义分析将原始的文本和图像特征映射到低维的概念空间(语义空间)得到文本和图像的语义特征;然后,通过语义特征训练分类器;最后,将分类器用于微博的情感分类。实验通过爬取的新浪微博数据集验证了本文提出的方法能够有效地提高多视图情感分类的效果,同时分析了几个常用特征的利弊。

1 潜在语义分析方法简介

1.1 潜在语义分析概念

在信息检索或者文本分析领域,通常使用向量空间模型[12](vector space model, VSM)来表示一篇文档。它将一篇文档或者一段话表示成向量,方便进行各种数学处理。虽然此种方法在一些应用中可以获得不错的效果,但在实际生活中,可能存在多次同义的问题,而 VSM 并不能很好地发现词与词之间在语义上的关系。

LSA 可以在一定程度上解决上述问题。LSA 源自信息检索领域问题:如何从 query 中找到相关的文档[13]。LSA 试图表达一个词背后隐藏的语义信息,它把词和文档都映射到一个语义空间并在这个空间内进行各种运算。这种想法是受到心理语言学家的启发[14]。LSA 认为文本中的词语存在着潜在的语义结构,同义词被映射到相同的语义空间之后应该有很大的关联度。

1.2 潜在语义分析

LSA 是一种无监督的学习技术,处理的是词-文档矩阵(在本文中处理的是文本和图像特征融合后的复数矩阵)。构建词-文档矩阵之后,LSA 通过使用奇异值分解[15](singular value decomposition, SVD)技术将词-文档矩阵分解,可以将原始高维空间中表示的词和文档投射到低维语义空间。

LSA 首先构造一个词-文档矩阵 $\boldsymbol{N}=[X_{ij}]$。其中矩阵的行表示词,列表示文档,X_{ij} 表示第 i 个词在第 j 个文档中的权重。矩阵中的一行 $\boldsymbol{t}_i=[x_{i1}\ x_{i2}\ \cdots\ x_{in}]$ 代表某个词和所有文档之间的关系,矩阵中的一列 $\boldsymbol{d}_j=[x_{1j}\ x_{2j}\ \cdots\ x_{mj}]^{\mathrm{T}}$ 代表某个文档和所有词语之间的关系($x_{ij}\neq 0$ 表示该文档包含该词汇,其值表示第 i 个词在第 j 个文档中的权重)。两个行向量的点积 $\boldsymbol{t}_i\cdot\boldsymbol{t}_p^{\mathrm{T}}$ 代表文档中两个词(第 i 个词和第 p 个词)的相关性;两个列的点积 $\boldsymbol{d}_j^{\mathrm{T}}\cdot\boldsymbol{d}_q$ 代表两个文档(第 j 篇文档和第 q 篇文档)的相关性。由于一个词一般出现在几个特定文档中,故矩阵 $\boldsymbol{N}$ 通常是一个稀疏矩阵。而通过奇异值分解,可以将高维的系数矩阵转化成低维的稠密矩阵。任何一个矩阵都可以使用奇异值分解[15],假设矩阵 $\boldsymbol{N}$ 为 $m\times n$ 矩阵,则奇异值分解定义如式(1):

$$\boldsymbol{N}=\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}} \tag{1}$$

式中:$\boldsymbol{U}$ 为 $m\times m$ 的矩阵,$\boldsymbol{\Sigma}$ 为 $m\times n$ 矩阵,$\boldsymbol{V}$ 为 $n\times n$ 矩阵。矩阵 $\boldsymbol{U}$、$\boldsymbol{V}$ 为奇异向量组成的正交方阵。$\boldsymbol{\Sigma}$ 是奇异值的对角矩阵,$\boldsymbol{\Sigma}=\mathrm{diag}(\sigma_1,\sigma_2,\cdots,\sigma_n)$,其中 σ_1,$\sigma_2,\cdots\sigma_n$ 是矩阵 $\boldsymbol{N}$ 的 n 个奇异值,且 $\sigma_1\geqslant\sigma_2\geqslant\cdots\geqslant\sigma_n$。得到了奇异值之后,取前 r 个最大的奇异值以及对应的特征向量即可以得到矩阵的低阶近似,如式(2)所示:

$$\boldsymbol{N}'\approx\boldsymbol{U}'\boldsymbol{\Sigma}'\boldsymbol{V}'^{\mathrm{T}} \tag{2}$$

式中:$\boldsymbol{U}'_{m\times r}=[u_1\ u_2\cdots\ u_r]$,$\boldsymbol{\Sigma}'_{r\times r}=\mathrm{diag}(\sigma_1,\sigma_2,\cdots,\sigma_r)$,$\boldsymbol{V}'^{\mathrm{T}}_{r\times n}=[v_1\ v_2\cdots\ v_r]^{\mathrm{T}}$,$\boldsymbol{\Sigma}'$为奇异值从大到小排列的对角矩阵,其中 r 的值远小于 m 和 n。目标是使得 $\boldsymbol{N}$ 与 $\boldsymbol{N}'$尽可能相似同时获得尽可能小的 r,其中 r 是语义空间的维度。之后,可以在该空间内计算词之间、文档之间以及词与文档之间的相似性度量等。

2 基于语义特征的多视图情感分类

现有的情感分类研究工作很多都是围绕文本展开的,但微博除了文本还存在大量的图片,如果能够将文本和图片结合,就可以获得比纯文本更多的信息量。但若仅仅使用原始特征,有可能带来维度灾难问题。

本文提出的基于语义特征的多视图情感分类方法将文本和图像特征并行融合,并使用 LSA 抽取

各自的语义特征,其流程如图 1 所示。

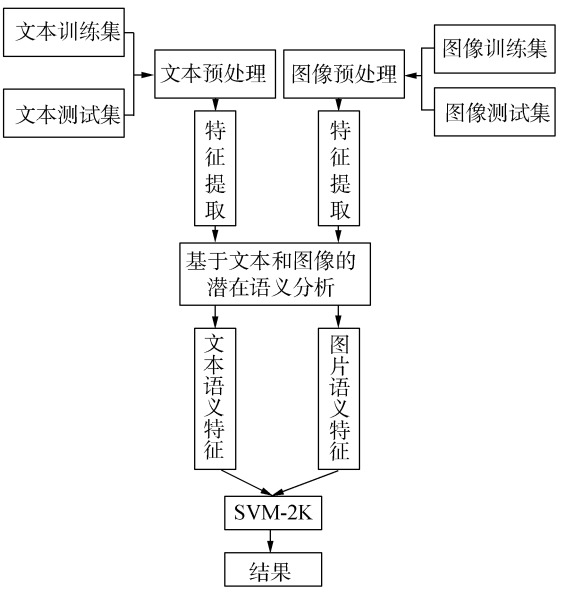


图 1　基于语义特征的多视图情感分类方法流程图

Fig.1　Flow chart of sentiment classification of microblogs based on semantic features

图像和文本的特征融合,在信息检索领域里的跨模型检索(cross-modal retrieval)中已经有相应的应用。但使用较多的融合方式是文本和图像特征的串行融合[10]。

Wang 等[16]在 3D 目标检索时,使用的两组特征串行融合方式如式(3)所示:

$$\boldsymbol{N}=\begin{bmatrix}\alpha_{11} & \alpha_{12} & \cdots & \alpha_{1n} & \cdots & \beta_{11} & \beta_{12} & \cdots & \beta_{1t}\\ \alpha_{21} & \alpha_{22} & \cdots & \alpha_{2n} & \cdots & \beta_{21} & \beta_{22} & \cdots & \beta_{2t}\\ \vdots & \vdots & & \vdots & & \vdots & \vdots & & \vdots\\ \alpha_{j1} & \alpha_{j2} & \cdots & \alpha_{jn} & \cdots & \beta_{j1} & \beta_{j2} & \cdots & \beta_{jt}\\ \vdots & \vdots & & \vdots & & \vdots & \vdots & & \vdots\\ \alpha_{m1} & \alpha_{m2} & \cdots & \alpha_{mn} & \cdots & \beta_{m1} & \beta_{m2} & \cdots & \beta_{mt}\end{bmatrix} \tag{3}$$

式中:m 表示样本的个数,n 表示第 1 个视图的特征维度,t 表示第 2 个视图的特征维度。α_{ij} 是第 i 个样本的第 j 维特征,β_{ij} 是第 i 个样本的第 j 维特征。

但是这样将两种不同属性的特征强行拼接在一个特征空间中,应用到微博中会失去原有的物理特性:一条微博是由文字和配图组成的整体。一条微博的文字和配图有一定的内在联系,而不是两个独立的个体。基于并行融合方法[10],本文对于融合前后的文本和图像特征使用复数进行表示。将文字图片的特征使用复数进行融合,可以反应微博的整体关系,即复数的实部表示文本特征,虚部表示图像特征。由于复数矩阵分解之后仍为复数矩阵,故分解之后的实部和虚部分别对应文本和图像的语义特征。

文本特征和图像特征融合方法如式(4)所示,将融合之后的复合特征称为一个新的文档 $\boldsymbol{d}_j$。

$$\boldsymbol{d}_j=\boldsymbol{\alpha}_j+\mathrm{i}\theta\boldsymbol{\beta}_j \tag{4}$$

式中:实部 $\boldsymbol{\alpha}_j$ 为文本特征向量,虚部 $\boldsymbol{\beta}_j$ 为图像特征向量,θ 是权重因子。极端情况:

当 $\theta\to 0$ 时,融合的特征 $\boldsymbol{d}_j\approx\boldsymbol{\alpha}_j$,此时近似于纯文本特征。

当 $\theta\to+\infty$ 时,则 $\boldsymbol{d}_j\approx\boldsymbol{\beta}_j$,即此时近似于使用纯图像特征的分类效果。

在本文工作中,我们将文本和图片同等对待,因此设 $\theta=1$。假设有 m 条微博,文本和图像的语义空间的维度为 n。那么由复数构成新的文档集合用矩阵表示如式(5)所示:

$$\boldsymbol{N}=\begin{bmatrix}\alpha_{11}+\mathrm{i}\beta_{11} & \alpha_{12}+\mathrm{i}\beta_{12} & \cdots & \alpha_{1n}+\mathrm{i}\beta_{1n}\\ \alpha_{21}+\mathrm{i}\beta_{21} & \alpha_{22}+\mathrm{i}\beta_{22} & \cdots & \alpha_{2n}+\mathrm{i}\beta_{2n}\\ \vdots & \vdots & & \vdots\\ \alpha_{j1}+\mathrm{i}\beta_{j1} & \alpha_{j2}+\mathrm{i}\beta_{j2} & \cdots & \alpha_{jn}+\mathrm{i}\beta_{jn}\\ \vdots & \vdots & & \vdots\\ \alpha_{m1}+\mathrm{i}\beta_{m1} & \alpha_{m2}+\mathrm{i}\beta_{m2} & \cdots & \alpha_{mn}+\mathrm{i}\beta_{mn}\end{bmatrix} \tag{5}$$

式中:α_{ij} 是第 i 条微博文本的第 j 维特征,β_{ij} 是第 i 条微博图像对应的第 j 维特征。

对上面的复数矩阵 $\boldsymbol{N}$ 进行奇异值分解并进行低阶近似,把高维的空间映射到低维的语义空间。将其映射到语义空间之后,再分别提取分解后低阶近似矩阵的每个元素的实部和虚部,得到文本和图片在低维空间的新特征,即语义特征。最后将提取的文本和图片的语义特征用于训练多视图分类器 SVM-2K[17],并使用测试集测试模型分类结果。具体步骤如下:

1)提取微博数据中的文本数据和图像数据,然后将文本和图像数据分成训练集和测试集。

2)分别对文本和图像进行预处理,并提取文本和图像的特征。

3)将文本特征和图像特征进行融合,形成一个复数矩阵。对该复数矩阵进行奇异值分解降维。将降维后的矩阵分离实部和虚部分别得到文本的

语义特征和图片的语义特征,语义特征提取过程如算法所示。

4) 将该语义特征在 SVM-2K 分类器中进行训练,然后用测试集验证。

5)得到测试集的情感分类结果。

算法 语义特征提取

输入 trainset, testset;

输出 lsa_trainset, lsa_testset。

1)txtimgtr←Text(trainset) + i×Image(trainset);

/* Text 函数取数据集中的文本数据,Image 函数取数据集中的图像数据,i 为虚数的单位 i */;

2)txtimage←Text(testset) + i × Image(testset);

3)COMPS_LSA←300 ;

4)comTxtImgTr← txtimgtr^{T}

/* txtimgtr^{T} 为矩阵 txtimgtr 的转置 */;

5)$[\boldsymbol{U},\boldsymbol{\Sigma},\boldsymbol{V}^{T}]$ = svd(comTxtImgTr,COMPS_LSA);

6)US←U(:, 1:COMPS_LSA);

7)SS← S(1:COMPS_LSA, 1:COMPS_LSA);

/* 对矩阵进行奇异值分解,取前 COMPS_LSA = 300 个最大的奇异值,也即为语义空间的维度 */;

8)comTxtImgTe←comTxtImgTe · U · inv(SS);

9)comTxtImgTr ←comTxtImgTr^{T} · US · inv(SS);

/* inv 为取矩阵的逆的函数 */;

10)lsa_Ttr←Text(comTxtImgTr);

11)lsa_Tte←Text(comTxtImgTe);

12)lsa_Itr←Image(comTxtImgTr);

13)lsa_Ite←Image(comTxtImgTe);

14)return lsa_Ttr+lsi_Itr, lsa_Tte + lsi_Ite。

3 实验

本节实验是为了验证多视图语义特征融合的有效性。我们使用了基于复数表示的文本特征和图像特征的并行融合方法,并将其进行潜在语义分析。将文本特征和图像特征分别映射到同维度下语义空间,得到各自的语义特征,将得到的语义特征用于训练分类器,最后使用测试集验证了微博情感分类的效果。

3.1 数据集

实验的数据集为爬虫从新浪微博的热门微博下爬取的。为了完成本文的任务,在爬取微博的时候仅仅保留同时含有文字和配图的微博。最终留下 1 000 条微博数据并手动进行标注。为了验证所提出方法的有效性我们采取了交叉验证的方式,其中 700 条数据作为训练集,300 条数据作为测试集。

数据采集过程如图 2 所示。

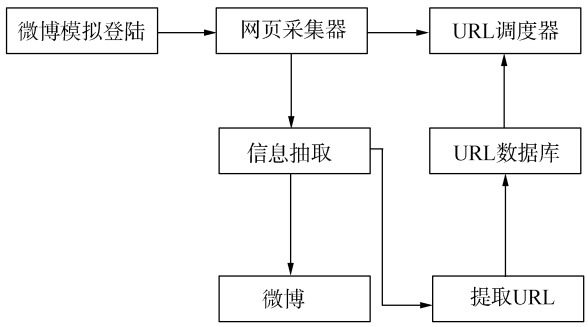


图 2 新浪微博数据采集过程

Fig.2 Sina micro-blog data acquisition process

将得到的微博数据作如下数据预处理:

1)过滤微博的一些冗余信息,如网址、转发对象、表情符号等。

2)将得到的微博文本和图像分离并编号,同一条微博的文本和图片编号相同。

3)分词:我们使用汉语分词系统 ICTCLAS[18] 对微博的文本进行分词。

4)去除停用词:分词后,去除一些无意义的停用词。

3.2 实验设计

实验中我们设置 LSA 语义空间的维度 $r = 300$,分别用向量空间模型(vector space model)和布尔模型(Bool model)进行加权。由于 Tan 等[19]已经证明对于情感分类来说,6 000 维度已经可以充分表示文本。除了选取 6 000 作为特征维度,我们展示了特征维度为 5 000 维下的实验结果。

在文本特征选择时,使用了文档频率(document frequency, DF)、互信息(mutual information, MI)、卡方分布(Chi-squared distribution, CHI)和信息增益(Information Gain, IG)这 4 种文本特征选择的方法,并比较了这 4 种特征做情感分类时的结果;对于图像,我们提取了图像的亮度、饱和度、色相、纹理、灰度共生矩阵。然后将提取的特征通过 LSA 映射到一个语义空间得到各自的语义特征,最后将文本和图像的语义特征使用 SVM-2K 进行分类,并使用测试集测试得到情感分类的结果。通过实验结果验证本文提出的基于文本和图像的语义特征情感分类方法的有效性。

3.3 实验结果

表 1 展示了文本特征为 5 000 维度时,使用纯

文本特征、纯图像与使用文本和图像结合的语义特征多视图分类的结果对比。表 1 对比了 DF、CHI、MI 和 IG 这 4 种文本特征选择方式对于不同分类方法结果的影响,表中的 SVM-2K 是指使用基于文本特征结合图像语义特征的多视图分类器。

表 1　5 000 维度的布尔模型

Table 1　5 000 dimensions of Bool model

特征提取方法	SVM 纯文本	SVM 纯图像	SVM-2K 文本+图像
DF	0.75	0.71	0.809
CHI	0.78	0.63	0.812
MI	0.745	0.653	0.806
IG	0.772	0.647	0.81
平均正确率	0.762	0.66	0.809

表 2 展示了文本特征为 6 000 维度时各种分类方法的对比,特征为布尔模型。

表 2　6 000 维度的布尔模型

Table 2　6 000 dimensions of Bool model

特征提取方法	SVM 纯文本	SVM 纯图像	SVM-2K 文本+图像
DF	0.742	0.623	0.791
CHI	0.763	0.658	0.795
MI	0.76	0.59	0.78
IG	0.77	0.61	0.77
平均正确率	0.759	0.620	0.784

表 3 展示了文本特征为 5 000 维度时,使用纯文本特征、纯图像与使用文本和图像结合的语义特征多视图分类的结果对比,同样对比了 DF、CHI、MI 和 IG 这 4 种特征选择方式对于各种分类方法结果的影响。

表 3　5 000 维度的向量空间模型

Table 3　5 000 dimensions of VSM

特征提取方法	SVM 纯文本	SVM 纯图像	SVM-2K 文本+图像
DF	0.62	0.53	0.65
CHI	0.78	0.69	0.81
MI	0.73	0.67	0.79
IG	0.72	0.65	0.806
平均正确率	0.712	0.635	0.764

表 4 展示了文本特征为 6 000 维度时各种分类方法的对比,特征的加权方式为向量空间模型。

表 4　6 000 维度的向量空间模型

Table 4　6 000 dimensions of VSM

特征提取方法	SVM 纯文本	SVM 纯图像	SVM-2K 文本+图像
DF	0.74	0.74	0.77
CHI	0.79	0.63	0.83
MI	0.72	0.62	0.82
IG	0.78	0.65	0.785
平均正确率	0.758	0.66	0.801

实验最后对比了不使用语义特征的多视图分类效果。为分析各个特征对于结果的影响,表 5 汇总了本文所提出方法情感分类精度结果。

表 5　基于语义特征的多视图情感分类方法分类精度统计

Table 5　Accuracy of multi-view sentiment classification of microblogs based on semantic features

特征提取方法	表 1	表 2	表 3	表 4	平均值
DF	0.809	0.791	0.65	0.77	0.755
CHI	0.812	0.81	0.81	0.83	0.816
MI	0.806	0.78	0.79	0.82	0.799
IG	0.81	0.77	0.806	0.785	0.793

3.4　实验分析

特征抽取方法的比较:通过表 5 可知,使用本文方法时 CHI 特征表现得最好,平均正确率为 81.6%;DF 表现得最不稳定,有时效果不错(如表 1 所示),有时表现得很差(如表 3 所示)。

语义特征:可以用不同的方式得到一个文档的语义特征,例如,可以用 LDA[20] 或者针对于文本较短的情况改进的 LDA 模型[21-22] 对文本进行聚类,用聚类的结果对文本进行再分析。图像也可以使用类似的方法。但把文本特征和图像特征分开进行语义映射,会失去二者的内在联系。

词项特征和语义特征:通过对比,我们可以发现,语义特征的分类精度最好的是 81.6%,最坏情况是 75.5%;而未经过 LSA 处理的纯文本特征最好情况是 75.75%,最坏情况是 74.5%。不难看出,使用经过 LSA 得到的语义特征,有助于提升微博情感分类的精度。不仅整体的分类效果更好,而且各个子分类器的分类效果也比纯文本特征有所提高。这表明,进行情感分类工作时在语义级别处理并行融合后特征能得到更好的分类效果。

在用户发的带有文本和图片的微博数据中,我们可以发现,本文所提出的基于语义特征的多视图微博情感分类方法的效果明显优于只考虑纯文本的情况。例如,微博“我希望躺在向日葵上,即使沮丧,也能朝着阳光”,其配图如图 3 所示。若使用纯文本将其分类得到的是负面的,而若采用本文提出的多视图语义特征方法将其分类得到的为正面情感。再如,微博“一个人不会,也不可能,将祂的全部呈现给你。你所看到的永远是祂的局部,而局部永远是美好的。”其配图如图 4 所示。若仅使用纯文本分类则分类结果为正面情感。采用本文提出的方法,则得到的是负面情感,而负面情感更加符合事实的判断。进而说明了本文方法的有效性。

图 3 示例 1 配图

Fig.3 Image in case 1

图 4 示例 2 配图

Fig.4 Image in case 2

4 结束语

本文首先利用并行特征融合方式,将文本和图像合理地组合在一起,然后用潜在语义分析技术,将文本和图像特征统一地映射到一个语义空间,最后使用多视图分类器 SVM-2K 进行分类。实验表明,基于本文多视图的语义特征方法的情感分类获得了比单纯的文本特征或者图像特征更好的效果。使用融合后的语义特征不管是文本特征做情感分类还是单从图像特征做情感分类,都比原来的分类精度有所提高。但是在 3.1 小节数据预处理时难免会剔除一些有用的信息,如表情、终端信息、转发信息、地理位置信息等。如何有效地利用这些因素提高情感分类精度有待进一步的研究。

参考文献:

[1] LIU B. Sentiment analysis and opinion mining[J]. Synthesis lectures on human language technologies, 2012, 5(1): 1-167.

[2] PANG T B, PANG B, LEE L. Thumbs up? Sentiment classification using machine learning[J]. Proceedings of EMNLP, 2002: 79-86.

[3] TÄCKSTRÖM O, MCDONALD R. Semi-supervised latent variable models for sentence-level sentiment analysis[C]// The 49th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, USA, 2011: 569-574.

[4] QIU G, LIU B, BU J, et al. Opinion word expansion and target extraction through double propagation[J]. Computational linguistics, 2011, 37(1): 9-27.

[5] WU Y, ZAHNG Q, HUANG X, et al. Phrase Dependency Parsing for Opinion Mining[C]//Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, USA, 2009: 1533-1541.

[6] LIU Y, HUANG X, AN A, et al. ARSA: a sentiment-aware model for predicting sales performance using blogs [C]//International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA, 2007: 607-614.

[7] MISHNE G, GLANCE N S. Predicting movie sales from blogger sentiment[C]//National Conference on Artificial Intelligence. Menlo Park, USA, 2006: 155-158.

[8] O'CONNOR B, BALASUBRAMANYAN R, ROUTLEDGE B R, et al. From tweets to polls: linking text sentiment to public opinion time series[C]//Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media. Menlo Park, USA, 2010: 122-129.

[9] CHIANG H C, MOSES R L, POTTER L C. Model-based Bayesian feature matching with application to synthetic aperture radar target recognition[J]. Pattern recognition, 2001, 34(8): 1539-1553.

[10] MCCULLOUGH C L. Feature and data-level fusion of infrared and visual images[J]. Proceedings of SPIE-the international society for optical engineering, 1999, 3719: 312-318.

[11] YANG J, YANG J Y, ZHANG D, et al. Feature fusion: parallel strategy vs. serial strategy[J]. Pattern recognition, 2003, 36(6): 1369-1381.

[12] SALTON G, WONG A, YANG C S. A vector space model

for automatic indexing[M]. New York: ACM, 1975: 613-620.

[13] DEERWESTER S, DUMAIS S T, FURNAS G W. Indexing by latent semantic analysis[J]. Journal of the american society for information science, 1990, 41: 391-407.

[14] REHDER B, SCHREINER M E, WOLFE M B W, et al. Using latent semantic analysis to assess knowledge: some technical considerations[J]. Discourse processes, 1998, 25(2/3): 337-354.

[15] GOLUB G H, REINSCH C. Singular value decomposition and least squares solutions[J]. Numerische mathematik, 1970, 14(5): 403-420.

[16] WANG F, PENG J, LI Y. Hypergraph based feature fusion for 3-D object retrieval[J]. Neurocomputing, 2015, 151: 612-619.

[17] FARQUHAR J D R, HARDOON D R, MENG H, et al. Two view learning: SVM-2K, theory and practice[C]// International Conference on Neural Information Processing Systems. Stroud sburg, USA, 2005: 355-362.

[18] ZHANG H P, YU H K, XIONG D Y, et al. HHMM-based Chinese lexical analyzer ICTCLAS[C]// Proceedings of the second SIGHAN workshop on Chinese language Processing-Volume 17. Stroudsburg, USA, 2003: 758-759.

[19] TAN S, ZHANG J. An empirical study of sentiment analysis for chinese documents[J]. Expert systems with applications, 2008, 34(4): 2622-2629.

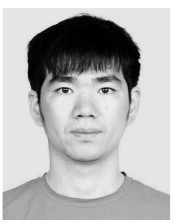
[20] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation[J]. Journal of machine learning research, 2003, 3: 993-1022.

[21] ZHAO W X, JIANG J, WENG J, et al. Comparing twitter and traditional media using topic models[J]. Lecture notes in computer science, 2011, 6611: 338-349.

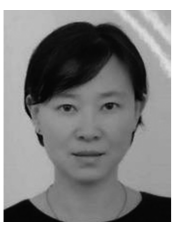
[22] YAN X, GUO J, LAN Y, et al. A biterm topic model for short texts[C]// Proceedings of the 22nd international conference on World Wide Web. New York, USA, 2013: 1445-1456.

作者简介:

吴钟强,男,1992 年生,硕士研究生,主要研究方向为文本挖掘、情感分析。



张耀文,男,1989 年生,硕士研究生,主要研究方向为文本挖掘、情感分析。



商琳,女,1973 年生,副教授,博士,主要研究方向为计算智能、机器学习、文本挖掘等。