第49卷第1期 中国科学技术大学学报 Vol. 49, No. 1
2019年1月 JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA Jan. 2019

文章编号:0253-2778(2019)01-0008-07

基于样本过滤和迁移学习的多领域情感分类模型

曲昭伟,赵燕娇,王晓茹

(北京邮电大学计算机学院,北京 100876)

摘要:目前,大部分进行情感分类的模型以单个数据集进行训练并测试,然而对一个数据集训练得到的模型参数不适用于另一个数据集,模型不具备通用性.为此提出一种适用于多个领域的情感分类模型(MDSC),借助样本过滤和迁移学习,使训练得到的模型参数适用于多个领域下的不同数据集,使模型更具适用性和拓展性,即先将文档映射到领域的分布式表示,并以此作为领域分类和情感分类的桥梁,最后进行情感分类.为了使模型更具通用性,需要选择代表性强的数据样本,于是通过构建具有领域独立性的情感字典对属于同一文档的句子进行过滤,获取高质量的训练集.同时为了提高分类准确率并减少训练时间,使用基于参数的迁移学习方法,利用神经网络获得文档向量再进行分类.在包含15个不同领域的数据集上进行实验,与其他情感分类模型相比得到了较好的实验效果.

关键词:情感分类;样本过滤;迁移学习;情感字典;神经网络

中图分类号:TP391 **文献标识码**:A doi:10.3969/j.issn.0253-2778.2019.01.002

引用格式:曲昭伟,赵燕娇,王晓茹.基于样本过滤和迁移学习的多领域情感分类模型[J].中国科学技术大学学报,2019,49(1):8-14.

英文作者,et al. A multi-domain sentiment classification model based on sample filtering and transfer learning[J]. Journal of University of Science and Technology of China, 2019,49(1):8-14.

A multi-domain sentiment classification model based on sample filtering and transfer learning

QU Zhaowei, ZHAO Yanjiao, WANG Xiaoru

(*School of Computer Science and Technology, Beijing University of Posts and Telecommunications, Beijing* 100876, *China*)

Abstract: Most of the models for sentiment classification are trained and tested on a single dataset. However, the model parameters obtained by training on one dataset are not suitable for another dataset and the model is not generic. A multi-domain sentiment classification model (MDSC) was proposed. With sample filtering and transfer learning, the trained model can be applied to different datasets in multiple domains and the model is more applicable and expandable. Specifically, a document is first mapped to the domain distribution which is used as a bridge between domain classification and sentiment classification, and then sentiment classification is completed. In order to make the model more generic, representative data samples should be selected. MDSC constructs a domain-independent sentiment lexicon to filter sentences that belong to the same document and obtain a high-quality training dataset. At the same time, to improve the classification accuracy and reduce the training time, parameter-based transfer learning with

收稿日期:2018-05-29;**修回日期**:2018-09-18

基金项目:国家自然科学基金(61672108)资助.

作者简介:曲昭伟,男,1970年生,博士/教授.研究方向:人工智能、数据挖掘、计算机网络技术.E-mail:zwqu@bupt.edu.cn

通讯作者:赵燕娇,硕士生.E-mail:zhaoyanjiao@bupt.edu.cn

neutral networks is used to obtain the document embeddings for classification. Extensive experiments on datasets containing 15 different domains show that the proposed model can achieve better performance compared with traditional models when applied to datasets in multiple domains.

Key words: sentiment classification; sample filtering; transfer learning; sentiment lexicon; neural network

0　引言

文本分类是自然语言处理(NLP)和文本挖掘领域的重要组成部分,其中情感分类作为文本分类的重要分支受到了研究者的广泛关注. 情感分类中的核心问题是特征表示,现在许多 NLP 系统通过使用一些通用表示作为文本的基本特征,节省训练时间的同时提升模型准确率,比如使用在大规模语料库上训练得到的词嵌入[1-2],之后有研究尝试获取大块文本如句子的通用表示[3],然而针对多领域通用情感分类模型的相关研究较少. 领域有几种定义,一般领域是一个包含不同实体的类[4],例如,电子产品、软件和 DVD 等在情感分类中视为不同的领域.

多数情感分类模型针对某一特定的数据集[5]进行设计或者适用于多个领域下的数据集[6],在一个数据集上训练得到的模型参数不能直接应用于其他数据集. 由于在属于相同领域的数据集中,影响其情感分类结果的因素类似,在属于不同领域的数据集中有差异[7],比如在电子产品域,“durable”和“light”代表积极情绪,“short battery life”代表消极情绪,而在书籍域,“exciting”和“thriller”代表积极情绪,“boring”和“lengthy”代表消极情绪,所以本文提出多领域情感分类模型,根据领域进行模型的训练,使其适用于多个领域下的数据集.

本文的主要贡献如下:

(Ⅰ)提出了一种适用于多个领域的情感分类模型. 借助领域分类将情感分类模型通用化,使训练得到的模型参数适用于多个领域的不同数据集,省去了对不同的数据集再次训练的时间,使模型更具适用性和拓展性.

(Ⅱ)通过构建具有领域独立性的情感字典,对数据集进行样本过滤,获取高质量的训练集,使模型更具有通用性.

(Ⅲ)使用基于参数的迁移学习方法,获取文档级别的特征,使其不偏向于某一领域,再用于分类,提高了分类准确率,并节省了训练时间.

1　相关工作

近期,自然语言处理中有大量针对情感分类的研究. 基本流程为先将输入向量化,再转换成句子或文档的向量表示,最后使用分类器进行情感分类. 许多研究通过改进神经网络结构进行句子或文档向量的提取,比如使用卷积神经网络(CNN)[8-10]或循环神经网络(RNN)[11-12],得到较好的分类效果.

准确地提取句子的语法和语义信息是语言理解的核心,文献[8]提出使用一种动态卷积神经网络进行句子建模,通过使用动态 k-max 池化可以处理不同长度的句子并捕捉长短关系,在分类和预测问题上取得了较好效果. 针对短文本包含的上下文信息有限进行情感分类时容易发生错误的现状,文献[9]提出了一种新的深度卷积神经网络,将字词向量组合进行句子级别的信息的获取,在二元分类的数据集上表现良好. 另外,文献[13]中使用包含多个组合函数的递归神经网络,根据输入向量自适应地选择组合函数,获取文本的矢量表示,该方法明显优于最新的情感分类方法. 上述研究的目的均为尽量提高情感分类的准确率. 另外还有些研究使用尽可能简单的方法,在保证准确率相对较高的同时减少训练时间,文献[1]表明,一个简单的句法无关的深度神经网络,在句法方差较高的数据集上的表现要好于句法模型,并且训练需要的时间更少. 以上研究均需在特定数据集上进行训练后进行评估,本文提出的模型 MDSC 可直接对多个领域的数据集进行分类,节省了训练时间.

2　基于样本过滤和迁移学习的多领域情感分类模型 MDSC

2.1　总体流程

我们使用符号 D 表示一个文档,一个文档包含多个句子 $\{s_i \mid i \in [1,N]\}$,每个句子 s_i 由一系列词 $W(s_i)=\{w_{ij} \mid j \in [1,M]\}$ 组成,使用 T 和 S 分别代表领域类别集合和情感类别集合,假设文档 D 的领域标签和情感标签分别为 t_D 和 y_D,其中 $t_D \in [1,|T|]$,$y_D \in \{-1,1\}$.

本文提出多领域情感分类模型(MDSC),模型的总体框架如图 1 所示. 模型主要包含两个模块:样本过滤模块和分类模块. 在样本过滤模块,我们利用

在多个领域中有标签的数据构建情感字典，提取出具有领域独立性的情感词，再根据情感词过滤出与文档情感极性一致的句子，从而获取高质量的数据集. 在分类模块，首先进行领域分类，再分别针对各领域训练情感分类模型，具体为使用基于参数的迁移学习方法获取文档向量，再利用多层神经网络预测文档的领域分布，最后根据得到的领域分布进行情感分类.

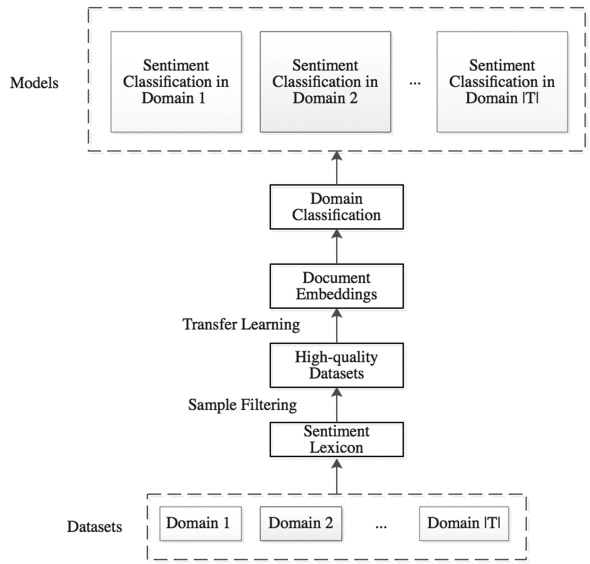

图 1 MDSC 的总体框架

Fig. 1 Overview of MDSC

2.2 样本过滤模型

分类器的性能在一定程度上与训练集的质量相关，一条高质量的训练样本意味着其所有的句子在情感极性上趋于一致，因此我们通过构建情感字典进行样本过滤，保留样本中与样本整体情感极性趋于一致的句子，删除与样本极性相反或者中立的句子，从而获得一个高质量的数据集.

首先进行情感字典的提取. 本文使用的数据集涉及多个领域，因此需要提取具有领域独立性的情感字典，使其可以在多个领域中进行情感极性的判定. 例如“good”是一个具有领域独立性的积极情感词，那么在任何领域中，包含该词的句子很有可能包含积极的情感. 我们采用加权对数似然比(WLLR)[14]来选择多个领域中积极和消极的词. 具体来说，首先在多个领域中标记句子的情感极性，并使用 NLTK 工具进行词性标注，然后提取出现频率大于 3 的形容词、副词和动词，删除否定词和停用词，接着，分别计算每个剩余的候选单词与正向情感和负向情感的相关程度，即

$$r(w,y)=p(w\mid y)\log\frac{p(w\mid y)}{p(w\mid y')} \tag{1}$$

式中，w 是一个单词，$y\in\{-1,1\}$ 是情感标签，y' 是与 y 相反的标签，$p(w\mid y)$ 是经验概率，即单词 w 在情感标签为 y 的句子中出现的概率. 分别按照 $r(w,-1)$ 值和 $r(w,1)$ 值对候选词进行降序排列. 最后，从两组排序的单词列表中分别选出前 25%作为具有正向和负向情感的单词，由此构成具有领域独立性的情感字典 L，$L=\{(w_i,r(w_i,y))\mid i\in[1,K]\}$.

接着对数据集进行样本过滤. 给定文档 D，将每个句子 s_i 中的单词与情感字典 L 进行映射，选出在情感字典 L 中出现的情感词，即 $W^L(s_i)=\{W^L_{ij_k}\mid k\in[1,M'],M'<M\}$，$W^L(s_i)\subseteq W(s_i)$，然后对情感词的分值求平均得到句子 s_i 的情感分数

$$\mathrm{score}(s_i)=\sum_{k=1}^{M'} r(W^L_{ij_k},y)/M' \tag{2}$$

式中，$r(W^L_{ij_k},y)$ 为情感词 $W^L_{ij_k}$ 在情感字典中的情感值. 单词的情感值和计算得到的句子情感分数均为实数，其中实数的正负代表极性，实数的绝对值代表情感程度. 选出与文档情感极性一致的句子，然后按分数的绝对值降序排列选出前 3 个句子(若少于 3 个则不做筛选)作为样本过滤后的文档，最终得到高质量的数据集.

在样本过滤模型中，因为训练数据集固定，即多个领域下的数据集，因此可以依据上述方法构建具有领域独立性的情感字典，该方法已经相对成熟，在实际应用中具备可操作性，并且通过一些手工检查表明，提取得到的情感词大多数确实是领域独立的. 因为该情感字典适用于多个领域，所以在之后的样本过滤操作中，情感字典不必重新提取，节省了时间.

2.3 分类模型

2.3.1 领域分类模型

首先进行文档向量的获取，即将文档映射到一个维度固定的分布式表示，该表示包含了文档的语法和语义信息. 为了使文档向量更具通用性，而不偏向于某一个领域，我们使用基于参数的迁移学习方法来获取文档向量[3]，具体为使用自然语言推理数(SNLI)[15]训练一个基于双向长短记忆网络(BiLSTM)[16]的编码模型，从而学习通用的文本表示. 其中 SNLI 数据集包含 570 k 人工生成的英文句子对，人工标注为以下 3 个类别之一：蕴含、矛盾

和中立，在该数据集上训练得到的编码模型能够抓取通用的语义信息，然后将此模型中进行文档向量提取的 BiLSTM[16] 模块，在参数保持不变的情况下迁移至本文的数据集，用于提取文档向量再进行领域划分，模型结构如图 2 所示. 给定文档 D 中的一系列单词 $\{\boldsymbol{w}_t\}_{t=1,\cdots H}$，BiLSTM 包含两个子网络，分别在不同的方向处理文本得到一组向量 $\{\boldsymbol{h}_t\}_{t=1,\cdots H}$ 如下：

$$\boldsymbol{h}_t^F = \overrightarrow{LSTM}(\boldsymbol{w}_1,\cdots,\boldsymbol{w}_H) \tag{3}$$

$$\boldsymbol{h}_t^B = \overleftarrow{LSTM}(\boldsymbol{w}_1,\cdots,\boldsymbol{w}_H) \tag{4}$$

$$\boldsymbol{h}_t = [\boldsymbol{h}_t^F,\boldsymbol{h}_t^B] \tag{5}$$

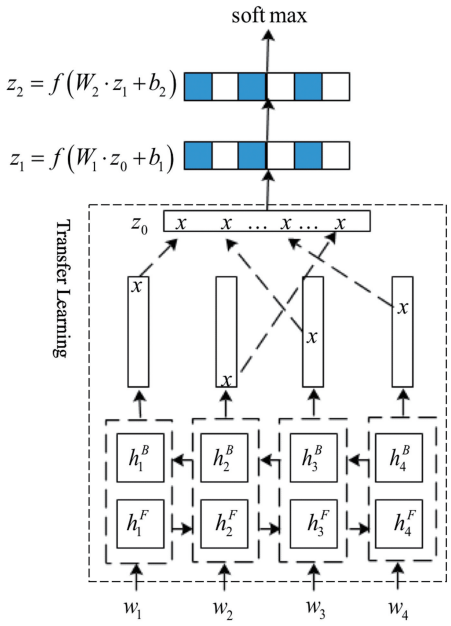

图 2　领域分类模型

Fig. 2　Domain Classification model

再经过最大池化层，通过筛选每一个维度的最大值将 $\{\boldsymbol{h}_t\}_{t=1,\cdots H}$ 合并得到一个维度为 d 的文档向量，即 $\boldsymbol{z}_0 = \max\{\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots\boldsymbol{h}_H,\}$，其中 max 函数是一个点对运算函数.

使用多层神经网络进行领域分类，假设神经网络的隐藏层有 n 层 $\boldsymbol{z}_{1\cdots n}^T$，我们计算每一层如下：

$$\boldsymbol{z}_i^T = f(\boldsymbol{W}_i^T \cdot \boldsymbol{z}_{i-1}^T + \boldsymbol{b}_i^T) \tag{6}$$

式中，$\boldsymbol{W}_i^T \in R^{d\times d}$ 是权重矩阵，$\boldsymbol{b}_i^T$ 是偏置项，f 是非线性转换函数. 然后将最后一层的向量表示(领域向量)输出到一个 softmax 分类器来预测领域标签，即

$$\boldsymbol{V}_T(D) = \text{softmax}(\boldsymbol{W}_c^T \cdot \boldsymbol{z}_n^T + \boldsymbol{b}_c^T) \tag{7}$$

式中，$\boldsymbol{W}_c^T$ 是一个 $|T|\times d$ 的矩阵，$\boldsymbol{b}_c^T$ 为偏置项，$\boldsymbol{V}_T(D)$ 是预测得到的领域概率分布.

2.3.2　情感分类模型

我们针对每一个领域进行情感分类模型的训练，即训练 $|T|$ 个情感分类模型，每一个情感分类模型和领域分类模型共享相同的文档向量 $\boldsymbol{z}_0$，然后将文档向量作为多层神经网络的输入进行情感分类，假设神经网络的隐藏层有 n 层 $\boldsymbol{z}_{1\cdots n}$，我们计算每一层如下：

$$\boldsymbol{z}_i^S = f(\boldsymbol{W}_i^S \cdot \boldsymbol{z}_{i-1}^S + \boldsymbol{b}_i^S) \tag{8}$$

式中，$\boldsymbol{W}_i^S \in R^{d\times d}$ 是权重矩阵，$\boldsymbol{b}_i^S$ 是偏置项，f 是非线性转换函数. 然后将最后一层的向量表示(情感向量)输出到一个 softmax 分类器来预测情感标签如下：

$$\boldsymbol{V}_S(D) = \text{softmax}(\boldsymbol{W}_c^S \cdot \boldsymbol{z}_n^S + \boldsymbol{b}_c^S) \tag{9}$$

式中，$\boldsymbol{W}_c^S$ 是一个 $|T|\times d$ 的矩阵，$\boldsymbol{b}_c^S$ 为偏置项，$\boldsymbol{V}_S(D)$ 是预测的情感极性的概率分布，最大值对应的情感极性即为预测结果.

我们希望得到适用于多个领域的情感分类模型，因此对于给定文档 D，首先使用领域分类模型得到领域的概率分布 $\boldsymbol{V}_T(D)$，然后经过属于 $|T|$ 个领域的情感分类模型，得到 $|T|$ 个情感极性的概率分布 $\{\boldsymbol{V}_S^i(D)\}_{i=1,\cdots T}$，将其根据 $\boldsymbol{V}_T(D)$ 进行加权得到 $\boldsymbol{V}_S^{\text{final}}(D)$，即

$$\boldsymbol{V}_S^{\text{final}}(D) = \sum_{i=1}^{|T|} \boldsymbol{V}_T^i(D)\,\boldsymbol{V}_S^i(D) \tag{10}$$

$\boldsymbol{V}_S^{\text{final}}(D)$ 会自动偏向于预测概率较大的领域，其中最大值对应的情感极性即为预测结果，从而得到给定文档 D 的情感类别.

2.3.3　模型训练

在多领域情感分类模型中，对于领域分类，我们使用带领域标签的数据并通过最小化交叉熵函数来训练模型. 给定一个训练样本 D，损失函数定义为：

$$L(D,t_D) = -\sum_{i=1}^{|T|} 1\{t_D = i\}\log V_T^i(D) \tag{11}$$

式中，$1\{t_D = i\} = 1$ 当且仅当 $t_D = i$ 成立. 对于情感分类，与领域分类类似，使用带有情感标签的数据通过最小化交叉熵损失函数进行模型的训练，给定一个文档 D，损失函数为

$$L(D,y_D) = -\sum_{i=-1,1} 1\{y_D = i\}\log V_S^i(D) \tag{12}$$

式中，$1\{y_D = i\} = 1$ 当且仅当 $y_D = i$ 成立. 我们利用 AdaGrad[17] 进行模型参数的更新.

3　实验

为了证明 MDSC 模型在多领域情感分类中的

适用性,在本节中我们将使用真实的开源数据集来进行模型验证实验并进行一些相关的讨论.

3.1 数据集

我们使用了 17 个不同的数据集,均用于二元情绪分类.其中前 14 个数据集是亚马逊产品评论数据[18],分别属于 14 个不同的领域.剩余的 3 个数据集为常用的电影评论数据集(IMDB[19]、MR[20]和 SST[21]).以上数据集均被划分为训练集、验证集和测试集,对应比例分别为 70%, 20%和 10%,具体数据统计如表 1 所示.

表 1 数据集详情

Tab.1 Statistics of datasets

Datasets	Train	Dev.	Test	Avg. L	Vocab.
Books	1 400	200	400	159	62 k
Electronics	1 398	200	400	101	30 k
DVD	1 400	200	400	173	69 k
Kitchen	1 400	200	400	89	28 k
Apparel	1 400	200	400	57	21 k
Camera	1 397	200	400	130	26 k
Health	1 400	200	400	81	26 k
Music	1 400	200	400	136	60 k
Toys	1 400	200	400	90	28 k
Video	1 400	200	400	156	57 k
Baby	1 300	200	400	104	26 k
Magazines	1 370	200	400	117	30 k
Software	1 315	200	400	129	26 k
Sports	1 400	200	400	94	30 k
IMDB	1 400	200	400	269	44 k
MR	1 400	200	400	21	12 k
SST	1 400	200	400	23	12 k

3.2 实验设置

在多领域情感分类模型中,我们使用反向传播算法进行模型的训练,词向量使用 300d GloVe[22]向量进行初始化,mini-batch 大小设为 32,使用 4 层神经网络,其中隐藏层维度为 300,其他参数通过从[−0.1,0.1]的均匀分布中随机采样进行初始化,并根据验证集上的性能来选择.

3.3 对比方法

将 MDSC 与其他情感分类模型对比,具体对比模型如下:

①LSTM[23]用于单个分类任务的标准 LSTM 模型;

②BiLSTM[16]用于单个分类任务的双向扩展 LSTM 模型;

③MSC[25]使用分类器融合的方法进行多领域的情感分类模型;

④MSC-WS[26]使用带有加权的分类器融合方法进行多领域的情感分类模型;

⑤ASP-MTL[6]使用对抗多任务学习进行多领域的文本分类模型.

对比模型均使用原始数据集进行训练,并且没有使用迁移学习获得文档向量.

3.4 实验结果和分析

首先,我们使用 MDSC 模型和以上对比模型在 15 个领域的数据集上进行训练并测试,在准确率方面的对比结果如表 2 所示.其中 LSTM 和 BiLSTM 模型是针对单个数据集(单个领域)的分类方法,因此分别在各领域的数据集上进行实验;MSC、MSC-WS 和 ASP-MTL 为多领域的分类模型,因此同时使用 15 个数据集进行模型的训练.从表 2 可以看出,MDSC 模型的平均准确率最高,并且在大多数领域达到最高准确率,可见该模型在多领域分类中的优势显著.其中多领域分类模型整体性能均优于单领域分类模型,主要因为多领域模型充分利用多领域下的数据集,通过集成各领域的信息,达到更高的准确率.

表 2 在所有领域上模型准确率对比

Tab.2 Results of models on all domains

Domains	LSTM	BiLSTM	MSC	MSC-WS	ASP-MTL	MDSC
Books	79.5	81.0	81.3	81.8	84.0	84.7
Electronics	80.5	78.5	81.5	83.5	86.8	87.4
DVD	81.7	80.5	80.4	81.0	85.5	87.8
Kitchen	78.0	81.2	78.2	79.5	86.2	85.9
Apparel	83.2	86.0	85.3	85.3	87.0	87.8
Camera	85.2	86.0	85.5	86.5	89.2	88.7
Health	84.5	78.7	86.3	87.8	88.2	89.6
Music	76.7	77.2	79.2	79.0	82.5	82.5
Toys	83.2	84.7	84.9	86.3	88.0	87.4
Video	81.5	83.7	83.1	83.3	84.5	87.9
Baby	84.7	83.5	85.6	86.3	88.6	88.2
Magazines	89.2	91.5	89.0	89.0	92.2	90.7
Software	84.7	85.7	84.5	85.5	87.2	89.9
Sports	81.7	82.0	82.7	82.5	84.7	85.9
Movies(SST)	84.9	86.5	87.1	87.5	89.0	89.2
Average	82.6	83.1	83.6	84.3	86.9	87.6

针对多领域模型对新数据集的适用性进行实验. 电影领域的数据集有多个,因为数据集的来源和结构存在不同,所以数据集之间存在一定的差异,我们选定其中一个数据集,并和其他 14 个领域的数据集结合用于模型的训练,然后在剩余的电影领域的数据集上进行测试,测试时模型的参数保持不变,实验结果如图 3 所示. 我们可以看出,当使用原有的模型参数进行新的数据集的情感分类时,MDSC 模型跟其他模型相比性能最好. 另外从图 3 可以看出,当使用 SST 数据集与其他领域的数据集结合进行模型训练时,在其他电影领域的数据集中准确率相对较高且比较稳定,因此选用 SST 作为电影领域的数据集进行后续模型的训练.

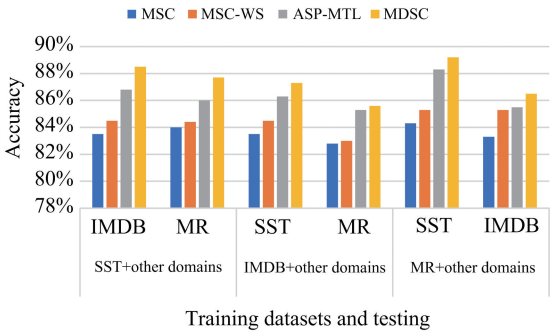

图 3　在新的数据集上模型准确率对比

Fig. 3　Results of models on new datasets

为了分析样本过滤和迁移学习对 MDSC 模型分类准确率的影响,我们使用 BASE、BASE-TRANSFER 和 MDSC 模型,在 15 个领域的数据集上进行实验并对比. 其中 BASE 模型代表在 MDSC 模型中使用原始数据集进行训练,未进行样本过滤,并且未使用迁移学习;BASE-TRANSFER 模型代表在 MDSC 模型中未进行样本过滤,但是使用迁移学习获取文档向量,实验结果图 4 所示. 从图 4 可以看出,MDSC 模型达到最高的分类准确率,BASE-TRANSFER 模型次之,BASE 模型的准确率最低,并且大体来说,BASE-TRANSFER 模型相对于 BASE 模型的增长量,比 MDSC 模型相对于 BASE-TRANSFER 模型的增长量多,可以得出迁移学习对 MDSC 模型的影响比样本过滤大. 由此可知,数据集的处理方式对模型的影响力不大,而通过迁移学习得到的文档向量直接和分类模型相关,对模型的预测概率的输出起主要作用.

我们通过进行相关实验来确定在多领域情感分类模型中神经网络的隐藏层层数,其中领域分类模型和情感分类模型的层数保持一致. 我们使用 15 个领域的数据集进行训练并测试,比较在不同层数的情况下,MDSC 与 BASE-TRANSFER 的平均准确率,实验结果如图 5 所示. 从图 5 可以看出,隐藏层数目为 2 或 3 的神经网络整体性能较好. 为了使得模型的参数相对较少,在 MDSC 模型中我们选用隐藏层为 2 层的神经网络.

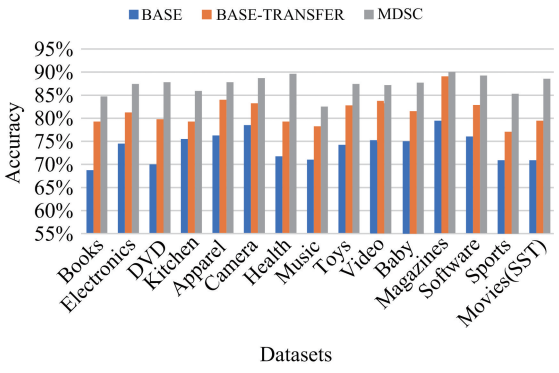

图 4　样本过滤和迁移学习的影响

Fig. 4　Effect of sample filtering and transfer learning

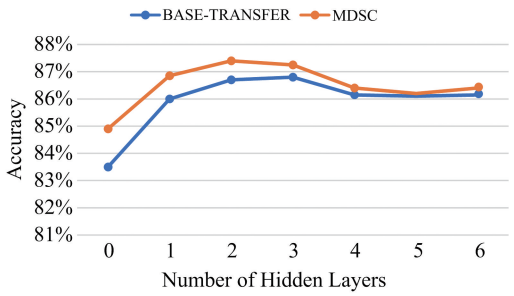

图 5　层数的影响

Fig. 5　Effect of layers

4　结论

本文提出了基于样本过滤和迁移学习的多领域情感分类模型(MDSC),使得情感分类模型适用于多个领域的数据集. 我们在包含 15 个不同领域的数据集上进行训练并测试,相比于其他情感分类模型,得到了较好的实验效果. 其中,构造具有领域独立性的情感字典进行样本过滤,并且使用迁移学习获取文档级别的特征,显著提高了 MDSC 模型的准确率. 在今后的工作中我们可以考虑跨模态进行信息的融合,通过将语音、图片或视频与文本进行结合,使得情感分类结果更准确,分类更详细. 还可以考虑将模型进行拓展,使其适用于跨领域的情感分析.

参考文献(References)

[1] IYYER M, MANJUNATHA V, BOYD-GRABER J, et al. Deep unordered composition rivals syntactic

methods for text classification[C]// Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Beijing: ACM Press, 2015, 1: 1681-1691.

[2] ZHOU P, QI Z, ZHENG S, et al. Text classification improved by integrating bidirectional LSTM with two-dimensional max pooling[J]. Computer Science, 2016, arXiv:1611.06639.

[3] CONNEAU A, KIELA D, SCHWENK H, et al. Supervised learning of universal sentence representations from natural language inference data [J]. Computer Science, 2017, arXiv:1705.02364.

[4] AL-MOSLMI T, OMAR N, ABDULLAH S, et al. Approaches to cross-domain sentiment analysis: A systematic literature review[J]. IEEE Access, 2017, 5: 16173-16192.

[5] REN Y, ZHANG Y, ZHANG M, et al. Context-sensitive twitter sentiment classification using neural network[C]// 13th AAAI Conference on Artificial Intelligence. Phoenix, USA: IEEE Press, 2016: 215-221.

[6] LIU P, QIU X, HUANG X. Adversarial multi-task learning for text classification[J]. Computer Science, 2017, arXiv:1704.05742.

[7] BOLLEGALA D, WEIR D, CARROLL J. Cross-domain sentiment classification using a sentiment sensitive thesaurus [J]. IEEE transactions on Knowledge and Data Engineering, 2013, 25(8): 1719-1731.

[8] KALCHBRENNER N, GREFENSTETTE E, BLUNSOM P. A convolutional neural network for modelling sentences[J]. 2014, arXiv:1404.2188.

[9] DOS SANTOS C, GATTI M. Deep convolutional neural networks for sentiment analysis of short texts [C]// Proceedings of the 25th International Conference on Computational Linguistics. Dublin, Ireland: IEEE Press, 2014: 9-78.

[10] COLLOBERT R, WESTON J, BOTTOU L, et al. Natural language processing (almost) from scratch[J]. Journal of Machine Learning Research, 2011, 12(12): 2493-2537.