doi: 10.3969/j.issn.0490-6756.2019.05.008

基于主题注意力层次记忆网络的文档情感建模

刘广峰, 黄贤英, 刘小洋, 范海波
(重庆理工大学计算机科学与工程学院, 重庆 400054)

摘 要: 针对文档水平情感分析传统模型存在先验知识依赖以及语义理解不足问题,提出一种基于注意力机制与层次网络特征表示的情感分析模型 TWE-ANN. 采用基于 CBOW 方式的 word2vec 模型针对语料训练词向量,减小词向量间的稀疏度,使用基于 Gibbs 采样的 LDA 算法计算出文档主题分布矩阵,继而通过层次 LSTM 神经网络获取更为完整的文本上下文信息从而提取出深度情感特征,将文档主题分布矩阵作为模型注意力机制提取文档特征,从而实现情感分类. 实验结果表明:提出的 TWE-ANN 模型较 TSA、HAN 模型分类效果较好,在 Yelp2015、IMDB、Amazon 数据集上的 F 值分别提升了 1.1%、0.3%、1.8%,在 Yelp2015 和 Amazon 数据集上的 RMSE 值分别提升了 1.3%、2.1%.

关键词: 文档分类; 情感分析; 层次记忆网络; 注意力机制; 词向量
中图分类号: TP391 **文献标识码**: A **文章编号**: 0490-6756(2019)05-0833-10

Document sentiment modeling based on topic attention hierarchy memory network

LIU Guang-Feng, *HUANG Xian-Ying*, *LIU Xiao-Yang*, *FAN Hai-Bo*
(College of Computer Science and Engineering, Chongqing University of Technology, Chongqing 400054, China)

Abstract: To the problem of prior knowledge and lack of semantic understanding in the traditional model of document level sentiment analysis, this paper proposes an sentiment analysis model called TWE-ANN (Attention Neural Networks based on Topic-enhanced Word Embedding), which is based on attention mechanism and hierarchical network feature representation. The word2vec model based on CBOW is used to train the word vector for corpus and the sparsity in the word vectors is reduced, the document topic distribution matrix is computed with LDA algorithm based on Gibbs sampling, the more complete text context information are obtained through hierarchical LSTM neural network and the deep sentiment features are finally extracted. The document topic distribution matrix is used as the model attention mechanism to extract the document features and the sentiment classification is thereby implemented. The experimental results show that the proposed TWE-ANN model has better classification results, compared with the TSA and HAN models. The F values on the Yelp2015, IMDB, and Amazon datasets is increased by 1.1%, 0.3%, and 1.8%, respectively, and the RMSE values on the Yelp2015 and Amazon datasets increased by 1.3% and 2.1%, respectively.

Keywords: Document classification; Sentiment analysis; Hierarchical memory network; Attention mechanism; Word embedding

收稿日期: 2018-10-30
基金项目: 国家社会科学基金 (17XXW004); 教育部基金 (16YJC860010); 2017 年重庆市教委人文社会科学研究项目(17SKG144); 2018 年重庆市科委技术创新与应用示范项目 (cstc2018jscx-msybX0049); 重庆市教委科学技术研究青年项目 (KJQN201801104); 重庆理工大学研究生创新项目 (ycx2018245)
作者简介: 刘广峰(1995-), 男, 山东滕州人, 硕士生, CCF 学生会员, 研究方向为自然语言处理、深度学习、情感分析.
通讯作者: 黄贤英. E-mail: wldsj_cqut@163.com

1 引 言

如今迅速发展的社交网络为用户提供了发表和分享个人言论的广阔平台,海量的评论数据由此而生.这些评论数据往往包含着个人情感取向,通过对其提取和分析可以促进社会上多个实际应用的发展,包括改善政府的舆情监控任务、为商家和消费者评估产品的市场价值等.如何利用自然语言处理技术来分析社交网络短文本的情感倾向,已经成为自然语言处理领域的一个热点研究方向[1].

情感分析,又称评论挖掘,它利用自然语言处理、文本分析、机器学习、计算语言学等方法对带有情感色彩的文本进行分析、处理、推理和归纳[2].其标准定义为:情感分析是对文本中关于某个实体的观点、情感、情绪及态度的计算研究[3].情感分析有两个核心任务:情感信息抽取以及情感信息分类,目前主要有基于规则、基于统计以及基于深度学习三种方法.针对基于规则的方法,文本中新词的出现、表达方式的变化以及复杂的语言处理等要素使得该方法难以适用.故目前流行使用基于统计和新兴的深度学习两种情感分析方法.如文献[4-6]通过人工设计的特征选择算法进行构建情感特征进而选择相对应的模型对训练集进行建模;Li 等[7]通过人工构建句子的标签和句子上下文标签特征来进行情感分类.这类基于统计的方法需要大量人工标注数据,一方面依赖于专业背景知识,另一方面未充分考虑到短文本的稀疏性.kim[8]将卷积神经网络用于文本领域进行情感分类,取得了较好的分类效果;Zhu 等[9]使用在图像视频领域表现优秀的 LSTM(Long Short-Term Memory)网络进行文本情感分析.这类基于深度学习的特征提取算法所取得的分类效果往往优于人工构建特征的方法,但在捕捉文本语义信息方面的能力仍然有所欠缺.

为解决情感分析中的数据稀疏性以及深层语义表达问题,word embedding 应运而生,这也导致许多 NLP 任务因为 word embedding 的加入取得了关键性突破.但直接使用词嵌入(word embedding)来进行情感分析往往容易忽略文档的主题情感信息,正如具有相反情感极性的词语(例如"好"和"坏")会被映射到同一类中,因为这两个词具有相似的用法和语法角色,这对于 NLP 中的词性标注任务来说是有很大帮助的,然而对于情感分析来说,由于它们具有相反的情感极性倾向,反而会降低情感分析模型的性能.为解决这一问题,已有许多学者将主题模型应用到情感分析领域.如 Phan 等[10]通过 LDA(Latent Dirichlet Allocation)模型从一个大型的 web 语料库中生成对应的隐藏主题并将其应用于短文本扩展.Vo 等[11]也是使用 LDA 模型对数据进行主题分析,是对各种类型的通用数据集进行建模继而达到文本特征扩展的目的.文献[12-13]在实验环节都验证了各自方法的有效性和准确性.

综上所述,现有情感分类模型强烈依赖于大量的先验知识且不能充分捕捉文本的语义信息.主题模型中的潜在狄利克雷分配模型正是为了捕捉文本潜在语义信息而提出的,具有良好的数学基础和灵活拓展性;词向量表示方法可以抽象的形式表示向量空间模型 (Vector Space Model,VSM),在一定程度上解决了背景依赖以及文本稀疏性问题;基于双向循环神经网络的层次表示模型可以充分地利用文本上下文信息以及潜在语义信息.因此,本文提出了结合主题信息与 Attention 机制的情感分析层次表示模型 TWE-ANN(Attention Neural Networks based on Topic-enhanced Word Embedding).

2 相关工作

2.1 情感分析的短文本特征扩展

大多数现有的短文本分类方法侧重于通过扩展短文本特征以克服数据稀疏性.特征扩展主要有两种方法:(1)通过获取外部文本来扩展短文本或者基于神经网络对数据进行增强;(2)对主题的潜在结构进行建模,以通过这些主题连接短文本.本文正是使用第二种方法来实现短文本特征扩展,且基于主题模型的方法已被大量运用于 NLP 领域并经实践证明效果显著.如 Kalchbrenner 等[12]提出动态的 CNN 模型,使用动态的池化层处理不同句长的句子,以此达到数据增强的目的.Cheng 和 Yan 等[13]为克服短文本中严重的数据稀疏性问题提出了 BTM(Biterm Topic Model)主题模型,通过对整个语料库的单词共现模式进行建模进而学习主题特征.Ren 等[14]通过递归自动编码器构建基于主题增强的词嵌入,进而将其作为情感分类特征输入到 SVM 分类器上进行情感极性判别,实验结果表明基于主题增强的词嵌入方式对于情感分类的性能提升极其有效.本文基于隐含狄利克雷分布 LDA 模型的 Gibbs 采样算法来对语料库文档生成主题特征.

2.2　Attention 机制与 LSTM 神经网络

近年来随着深度学习在计算机视觉、语音识别领域的成功应用，基于深度学习的模型越来越成为自然语言处理领域中的情感分析的主流方法. 在自然语言处理领域中，文本往往具有时序信息，在获取文本语义特征时结合了时序特征往往会在一定程度上促进分类性能的提升，而循环神经网络中的 LSTM 对于文本的时序特征和语义特征都可以兼顾. 如刘全等[15]基于 CNN 和 LSTM 网络的融合提出一种情感分析深度模型. 支淑婷等[16]提出融合多注意力和属性上下文的长短时记忆神经网络模型. 文献[15-16]所提出的模型在实验中取得的分类效果都明显优于基于传统机器学习的方法. 然而将主题模型应用到情感分析领域时，每一篇文档所具有的主题分布是不同的，直接将主题特征与通过深度学习得到的文档特征拼接会忽略一些更为重要的情感信息，而 Attention 机制正是用来解决这个问题而出现的. Attention 机制本是模拟人类注意力运行机制，最初是用在图像处理领域，目的是为了让神经网络在处理数据时重点关注某些信息. Yang 等[17]基于 Attention 机制与 GRU(Gated Recurrent Unit)网络提出了一个分层注意力网络用于文档分类，充分发挥了 Attention 机制的优越性. 胡朝举等[18]将 Attention 机制和 LSTM 结合解决特定主题的情感分析任务. 本文将通过 LDA 生成的主题模型作为 Attention 机制，在模型训练时赋予各个文档的主题权重分布信息.

2.3　CBOW 模型

word2vec 模型是由 Tomas 等在 Log-Bilinea 和 NNLM 两个模型的基础上开发的工具，可以将词从高维空间分布式地映射到低维空间且保留了词向量之间的位置关系，从而解决了向量稀疏和语义联系两个问题，其分为 CBOW(Continuous Bag-of-Words)和 Skip-gram 两种方式. CBOW 模型的训练输入某一个特征词的上下文相关的词对应的词向量，而输出的是这个特征词的词向量. Skip-Gram 模型和 CBOW 的思路相反，即输入一个特征词的词向量，而输出的是这个特征词对应的上下文词向量. 本文采用 CBOW 模型来训练词向量.

在 CBOW 模型中，给定词 k 对应的上下文向量 context(k)，根据 context(k)去预测 k. 若指定 context(k)，k 是一个正样本，其余词作为负样本，于是可以通过负采样得到关于 k 的负样本集 Neg(k). 词的特征可按照如下表示.

$$F^k(\bar{k})=\begin{cases}1, & \bar{k}=k\\ 0, & \bar{k}\neq k\end{cases} \tag{1}$$

CBOW 模型目标函数可按照如下表示.

$$J(k)=\prod_{c\in(k)\cup Neg(k)} P(c\,|\,\mathrm{context}(k)) \tag{2}$$

其中，

$$P(c\,|\,\mathrm{context}(k))=\begin{cases}\sigma(X_k^T\theta^c), & F^k(c)=1\\ 1-\sigma(X_k^T\theta^c), & F^k(c)=0\end{cases} \tag{3}$$

3　提出的 TWE-ANN 模型

对比已有的基于短文本特征扩展的情感分析模型，本文提出了基于主题增强与 Attention 机制的 TWE-ANN 情感分析模型. 使用 LDA 模型中的 Gibbs 采样算法以获得文档主题分布矩阵，再将此矩阵作为提出模型的 Attention 机制，模型结构如图 1 所示.

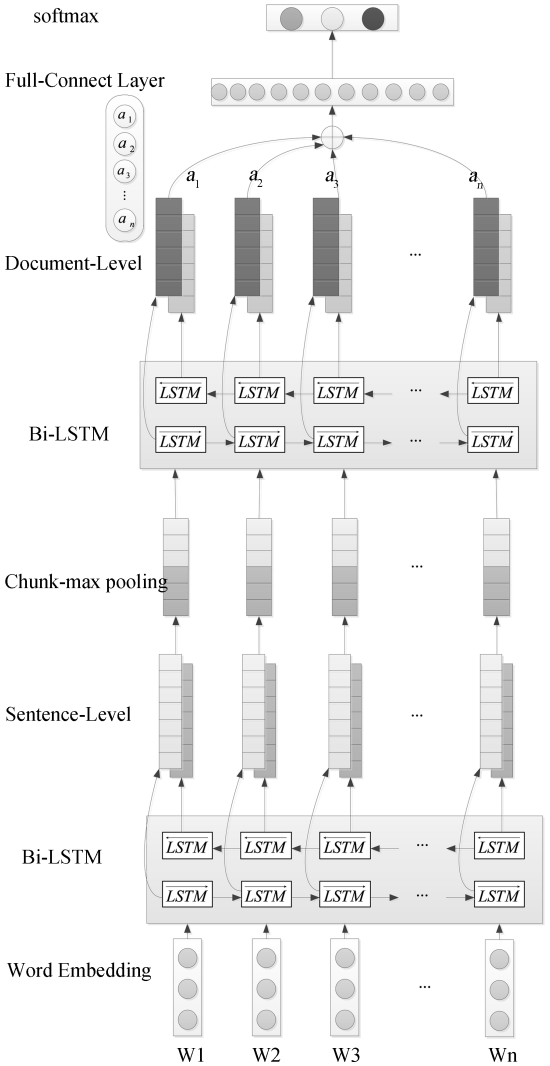

图 1　TWE-ANN 情感分析模型

Fig. 1　TWE-ANN sentiment analysis model

实验主要流程为:首先将基于 word2vec 的 CBOW 模型训练得到的词向量作为第一层 LSTM 网络的输入,将得到的隐藏层经过聚合处理得到句子向量;再对句子向量进行 chunk-max pooling 处理生成带有位置信息的句子向量;然后将句子向量通过第二层 LSTM 网络得到文档隐藏层;再将得到的文档隐藏层结合 Attention 机制得到全连接层特征;最后利用 softmax 层完成情感分析工作.

3.1 基于 LSTM 的序列编码

自然语言处理领域中的文本数据往往都包含着时序信息,这些信息在解决自然语言处理领域的许多任务时都发挥了极其重要的作用,而作为通过链式神经网络结构传播历史信息的循环神经网络(Recurrent Neural Network,RNN)正是用来处理时序数据的经典网络. 在处理时序数据时,当前输入 x_t 以及上一时刻输出的隐藏状态 h_{t-1} 对于 RNN 是透明可见的. 然而对 RNN 进行训练时,发现其内部由于梯度消失/爆炸导致不能对长期依赖进行针对性处理[19]. 为了弥补 RNN 对于长期依赖的不足,LSTM 网络应运而生,其自然行为便是长期的保存输入,并在自然语言处理领域表现出很好的应用效果.

因此,本文使用 LSTM 对文本信息进行序列编码. LSTM 的优势在于其具有三种特殊的门函数:输入门、遗忘门和输出门,通过这三种门来控制神经网络的记忆. 如图 2 所示,单个 LSTM 记忆单元在某一时刻 t 的前向计算过程如下[20].

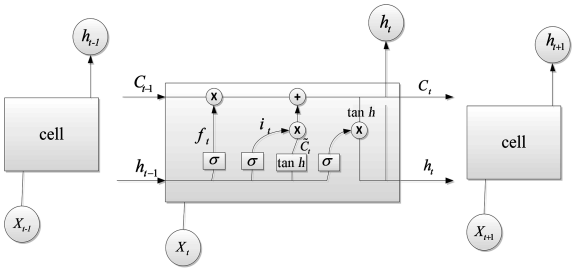

图 2 LSTM 单元内部架构

Fig. 2 LSTM unit internal architecture

(1) 遗忘门机制:

$$f_t=\sigma(W_f\cdot[h_{t-1},x_t]+b_f) \tag{4}$$

(2) 输入门机制:

$$i_t=\sigma(W_i\cdot[h_{t-1},x_t]+b_i) \tag{5}$$

$$\widetilde{C}_t=\tan h(W_C\cdot[h_{t-1},x_t]+b_C) \tag{6}$$

$$C_t=f_t\times C_{t-1}+i_t\times\widetilde{C}_t \tag{7}$$

(3) 输出门机制:

$$o_t=\sigma(W_o\cdot[h_{t-1},x_t]+b_o) \tag{8}$$

$$h_t=o_t\times\tan h(C_t) \tag{9}$$

式(4)~(9)中,$\{W_*,b_*\}$是神经网络训练的参数集合;$\widetilde{C}_t,f_t,i_t,o_t$ 分别表示时刻 t 记忆单元的输入单元、遗忘门、输入门和输出门的输出值;h_{t-1},x_t 分别表示时刻 t 上一个记忆单元以及当前的记忆单元的输入;C_t 表示时刻 t 记忆单元的内部状态;h_t 表示时刻 t 记忆单元的输出.

3.2 基于主题增强的注意力机制

主题模型已被成功应用于许多情感分析模型中[12-14],主要分为基于 Gibbs 采样的 LDA 算法和基于变分推断 EM 的 LDA 算法. 本文使用 Gibbs 采样的 LDA 算法对语料库中的文档集合进行主题建模,以获得文档对应的主题矩阵.

3.2.1 *文档主题分布矩阵计算* 本文希望通过 LDA 算法获得文档主题分布矩阵作为情感分析模型的注意力机制,文中矩阵计算步骤如下.

(1) 选择合适的主题数 k,选择合适的超参数向量 $\vec{\alpha},\vec{\eta}$;(2) 对应语料库中每一篇文档的每一个词,随机的赋予一个主题编号 Z;(3) 重新扫描语料库,对于每一个词,利用 Gibbs 采样公式更新其主题编号,并更新语料库中该词的编号;(4) 重复第 2 步的基于坐标轴轮换的 Gibbs 采样,直到 Gibbs 采样收敛;(5) 统计语料库中的各个文档各个词的主题,得到文档主题分布 θ_s.

对于每一个词 W,本文根据多项式分配公式(10)对该词的主题分配进行 Gibbs 采样.

$$P(Z_W=K|\vec{Z}_{-W},\vec{W},\alpha,\beta)=\frac{n_{K,-W}+\beta}{\sum_{v=1}^{|\vec{W}|}(n_{K,v}+\beta)-1}\cdot\frac{n_{d,-W}^{K}+\alpha}{\sum_{j=1}^{K}(n_d^j+\alpha)-1} \tag{10}$$

其中,$\vec{Z}_{-W}$表示除当前分配之外的所有词语的主题分配;$n_{k,-W}$表示除当前分配之外的主题 K 分配给词 W 的次数;$\sum_{v=1}^{|\vec{W}|}n_{k,v}-1$ 表示除当前分配外的主题 K 分配给词汇表中所有词的总次数;$n_{d,-W}^{K}$表示除当前分配外,分配给主题 K 的文档 d 中的词语数量;$\sum_{j=1}^{K}n_d^j-1$ 表示除当前词语外,文档 d 中的词语总数. 在 Gibbs 采样的最后一次迭代中,保存每个单词的主题分配以用于扩展语料库和进一步的主题/词向量学习.

3.2.2 *文档主题注意力* 并非所有单词对句子情感信息的表达起到同等作用,同理,句子对于文档的意义也会有所不同,因此,我们引入注意力机制并尝试结合文档主题分布矩阵来提取对文档情感信息表达起到重要作用的主题特征,进而将其表示

进行汇总. 具体计算方式如式(11)～(13)所示.

$$u_i = \tan h(W_s h_i + b_s) \tag{11}$$

$$\alpha_i = \frac{\exp(u_i^T \cdot \theta_s)}{\sum_i \exp(u_i^T \cdot \theta_s)} \tag{12}$$

$$f = \sum_i \alpha_i \cdot h_i \tag{13}$$

其中，W_s 代表权重矩阵；b_s 表示偏置项；u_i 类似于打分函数，用于衡量文档主题特征的重要程度，继而通过 softmax 函数计算文档主题向量 θ_s 的权重值 α_i. 最后将 α_i 与对应的文档水平的隐状态向量进行加权求和得到向量 f.

3.3　层次特征表示模型

如图 1 所示，本文构建的 TWE-ANN 模型由 Word Embedding Layer，Bi-LSTM Layer，Chunk-Max Pooling Layer，Bi-LSTM Layer，Fully-Connect Layer 以及 Softmax Layer 组成，每一层的输出为下一层的输入.

利用该模型首先获取句子水平层面的情感特征，继而获取文档水平层面的情感特征，将其用于情感分类和判别. 模型构造方法如下.

(1) Word Embedding Layer. 该层为模型的输入部分，即将语料库中的一段文本 T 输入进行词嵌入处理.

本文采用谷歌开源工具 Word2Vec 基于 CBOW 模型对语料进行训练从而获得文本词向量表示. 词向量可捕捉从语料中的词语到实数维向量空间的复杂映射，指定词向量空间为 Ψ，其大小为 $|\Psi| \times m$，Ψ 中每一行表示某个单词的 m 维词向量，$|\Psi|$ 表示词向量中包含词语的个数. 语料库中的一条评论文本 T 可表示为如下序列

$$(t_1, t_2, \cdots, t_n) \tag{14}$$

其中，n 表示文本 T 中的词语个数；t_i 表示 T 中第 i 个词语($1 \leqslant i \leqslant n$).

若将 T 转换为词向量矩阵，首先应在 Ψ 中搜索词 t_i 对应的词向量，若存在则选中对应的词向量，用 W_i 表示，否则将对应的词向量即 W_i 置为 0. 在找到每一个词语对应的词向量之后，将每一个词向量堆叠形成词向量特征矩阵 W，其大小为 $n \times m$，W 的每一行表示语料库中一个词语对应的词向量，可表示如下.

$$(t_1, t_2, \cdots, t_n) \Rightarrow (W_1, W_2, \cdots, W_n)^{\mathrm{T}} \tag{15}$$

(2) Bi-LSTM Layer. LSTM 单元将文本中的词语作为输入，经过处理后产生与该单词对应的隐藏状态输出 $H = (h_1, h_2, \cdots, h_n)$，其中 h_i 是 LSTM 在第 i 个时间步的隐藏状态信息，包括句子中从开始词语 t_1 到该词语 t_i 的所有信息. 本文通过构造两个 LSTM 神经网络来实现从两个相反的方向获取信息，更有利于从整体上捕捉句子的长依赖关系以及文本的深层语义表达，两个神经网络的输入一致.

基于 3.1 节介绍的 LSTM 结构，该层操作为

$$\overrightarrow{h}_i = \overrightarrow{LSTM}(W_i, \overrightarrow{h}_{i-1}) \tag{16}$$

$$\overleftarrow{h}_i = \overleftarrow{LSTM}(W_i, \overleftarrow{h}_{i+1}) \tag{17}$$

$$h_i = \overrightarrow{h}_i \,\|\, \overleftarrow{h}_i, h_i \in R^{2L} \tag{18}$$

其中，$\overrightarrow{h}_i$、$\overleftarrow{h}_i$ 分别表示前向 LSTM 以及反向 LSTM 在时刻 i 的输出向量；$\|$ 表示连接操作；L 表示每一个 LSTM 的单元数量. 经过以上处理后即生成了句子水平的特征向量.

(3) Chunk-Max Pooling Layer. 针对时序数据中的特征位置信息以及特征强度，文献[21-22]证明了在进行池化操作时结合上述信息可有效促进模型性能的提升. 故本文采用 Chunk-Max Pooling 针对句子水平的特征向量进行池化操作以获取向量中显著的特征值，同时在一定程度上捕获了粗粒度的特征位置信息以及特征强度. 计算方式如下所示. 具体操作如图 3 所示.

$$\bar{h}_i = \max\{h_i(m)\} \tag{19}$$

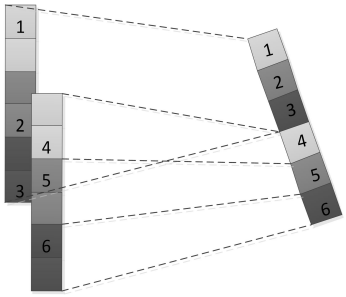

图 3　Chunk-Max Pooling 示意图
Fig. 3　Chunk-Max Pooling schematic

图 3 中，选择池化单元的宽度及其高度均为 2，步长同样设置为 2，经过池化处理后，原来的两个 feature map 就整合成了一个 feature map，在一定程度上也达到了降维的目的. 至此就完成了文档句子水平层面的特征提取工作.

(4) Bi-LSTM Layer. 上述过程描述了使用双向 LSTM 针对文档句子水平层面的特征提取过程，此处针对得到的特征继续使用双向 LSTM 进行处理从而获得文档水平层面的特征. 具体操作为

$$\overrightarrow{hd}_i = \overrightarrow{LSTM}(W_i, \bar{h}_{i-1}) \tag{20}$$

$$\overleftarrow{hd}_i = \overleftarrow{LSTM}(W_i, \bar{h}_{i+1}) \tag{21}$$

$$hd_i = \overrightarrow{hd}_i \,||\, \overleftarrow{hd}_i \tag{22}$$

(5) Full-Connect Layer. 该层基于 3.2 节介绍的基于主题增强的注意力机制对文档水平层面的特征

向量 hd_i 以及文档主题分布矩阵 θ_s 进行操作.

$$u_i=\tan h(W_s h_{di}+b_s) \tag{23}$$

$$\alpha_i=\frac{\exp(u_i^T\cdot\theta_s)}{\sum_i\exp(u_i^T\cdot\theta_s)} \tag{24}$$

$$f=\sum_i\alpha_i\cdot h_{di} \tag{25}$$

其中,f 即为生成的全连接层向量,用于接下来的文档情感分类.

3.4 特征规范化

本文在模型的 word embedding 层添加了高斯噪声,它是一种随机的数据增强技术,可以使模型更好的避免过拟合的出现. 此外,本文使用了 dropout 来随机丢弃网络中的神经元. Dropout 可防止网络神经元出现共同适应性,对于每一个训练样本,dropout 使整个网络中的某个子网络参与训练,因此它也被认为是一种集成学习的形式. 本文还将 dropout 应用到了双向 LSTM 网络中.

3.5 文档情感分类

使用 TWE-ANN 模型对文档进行情感分析. 全连接层向量 f 是文档的高级表示,可用作文档分类的情感特征.

$$p=\text{softmax}(W_f f+b_f) \tag{26}$$

使用交叉熵损失函数作为优化目标函数,使用反向传播算法计算并迭代更新模型参数如下.

$$L=-\sum_{d\in D}\sum_{s=1} p_s^c\cdot\log(p_s(d)) \tag{27}$$

其中,D 代表训练集中的文档集合;s 代表情感类别;p_s^c 代表情感为 s 的 0−1 分布. 当文档情感类别为 s 时,p_s^c 取值为 1,否则取值为 0. $p_s(d)$表示文档 d 的情感类别是 s 时的概率大小.

4 实验结果分析

4.1 实验数据集

在 3 个大规模文档分类数据集上评估 TWE-AGNN 模型的有效性,各个数据集的统计信息如表 1 所示. 本文规定:将 80% 的数据用于训练,10%的数据用于验证,剩余 10%的数据用于测试.

Yelp2015 数据集:评论文本来自 2015 年的 Yelp 评论挑战赛,与 Tang 等[23]使用的 Yelp2015 数据集信息一致,其中评论级别共有 5 个:1~5,级别越高越好;IMDB 数据集:来自 Diao 等[24]使用的数据集. 其中评论级别共有 10 个:1~10,级别越高越好;Amazon 数据集:来自 Zhang 等[25]使用的数据集. 其中评论级别共有 5 个:1~5,级别越高越好.

如表 1,单词最大数量和单词平均数量代表在一个数据集中对每一篇文档进行统计得到的数值,同理,句子最大数量和句子平均数量也是如此统计.

表 1 数据集统计信息

Tab. 1 Data set statistics

信息	Yelp2015	IMDB	Amazon
类别数目	5	10	5
文档数目	1 569 264	348 415	3 650 000
单词最大数量	1 199	2 802	596
单词平均数量	151.9	325.6	91.9
句子最大数量	151	148	99
句子平均数量	9.0	14.0	4.9
词汇表大小	612 636	115 831	1 919 336

4.2 评价指标

本文选取 F 值、$RMSE$(平方根误差)作为评价指标. 其中,F 值用来衡量分类的整体效果,平方根误差用来衡量预测值和真值之间的离散程度.

如表 2 所示,用 P 表示对测试集进行情感分类后,预测为某个类别的样本中真正类别的样本所占的比例,即

$$P=\frac{r}{r+t} \tag{28}$$

如表 2 所示,用 R 表示对测试集进行情感分类后,预测为某个类别中的真实类别占所有真实类别的比例,即

$$R=\frac{r}{r+s} \tag{29}$$

表 2 分类判别混淆矩阵

Tab. 2 Classification discriminant matrix

真实结果	预测结果	
	属于类别 C	不属于类别 C
属于类别 C	r	s
不属于类别 C	t	z

为了对准确率 P 和召回率 R 进行综合考虑,本文使用两者的加权调和平均数 F 来衡量最终分类效果.

$$F=\frac{2\times P\times R}{P+R} \tag{30}$$

对于平方根误差($RMSE$)可以根据以下公式来计算.

$$RMSE=\sqrt{\frac{\sum_i(\bar{y}-y)^2}{N}} \tag{31}$$

其中,N 表示文档总数;$\bar{y}$表示预测值;y 表示真实值.

4.3 实验参数设置

本文在读取文档时,将每一篇文档切分为句子

集合然后使用斯坦福的 CoreNLP 工具[26]对每一个句子进行标记.

在构建词汇表时，本文采用单因子变量实验确定最佳词频阈值(如图 4 所示)，并对低于词频阈值的词语采用‘UNK’这一特殊字符替换. 本文通过在数据集上执行 word2vec 模型的无监督式训练从而获得词嵌入，然后使用 word embedding 来初始化词向量空间 Ψ.

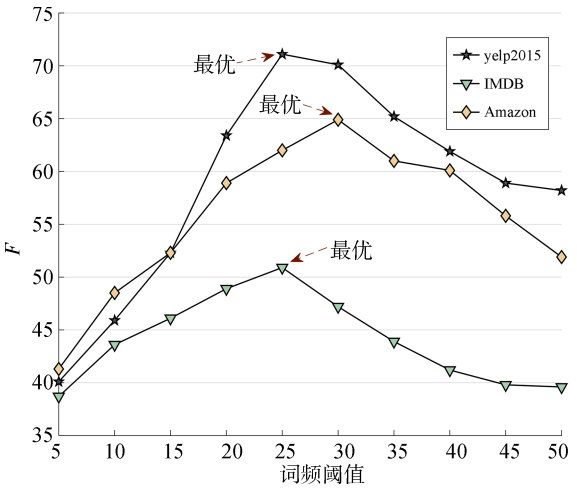

图 4　基于词频阈值的 F 值趋势变化

Fig. 4　Trend change of F value based on word frequency threshold

本文通过观察模型在验证集上的表现来对模型的超参数进行微调. 经过单因子变量实验后(如图 5 所示)，本文将词嵌入维度设置为 200，将 LSTM 网络的单元个数设置为 50，双向 LSTM 神经网络生成的向量维数为 100，因此，本文生成的文档主题分布矩阵 θ_s 的维数也设置为 100.

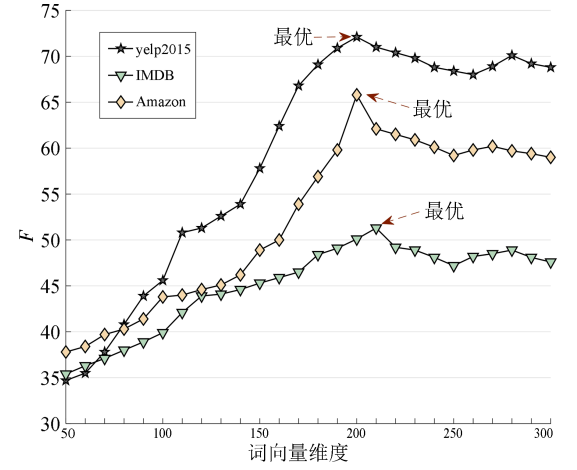

图 5　基于词向量维度的 F 值趋势变化

Fig. 5　Trend change of F value based on word vector dimension

对于模型训练，本文将批量大小设置为 64，Chunk-Max Pooling 的段数设置为 3，使用动量为 0.9 的 SGD(Stochastic Gradient Descent，随即梯度下降)来训练所有的模型，通过在验证集上进行网格搜索来确定最佳学习率.

表 3　实验参数设置

Tab. 3　Experimental parameter setting

参数	参数/函数名称	参数值/函数值
——	激活函数	tanh
——	优化方法	SGD
——	dropout	0.5
lr	模型学习率	0.0003
m	词向量维度	需对比实验[50−300]
D	词频阈值	需对比实验[5−50]
h	LSTM 的单元数量	50

如图 4 所示，随着词频阈值从最初的 5 渐渐增大，F 值也在不断提升，F 值在 Yelp201、IMDB 以及 Amazon 数据集分别达到最优的词频阈值是 25、25、30，进而随着词频阈值继续增加，F 值在各自数据集上反而逐渐减小. 因此，在神经网络训练时，选择这三者最优值出现次数较多的 25 作为参数词频阈值 D 的值.

如图 5 所示，参数词向量维度 m 初值为 50，随着 m 不断增大，F 值也在缓慢提升，直到 m 达到 200 时，F 值在 Yelp2015 和 Amazon 数据集上达到最优，而在 IMDB 数据集上的最优值是 210. 随后尽管 m 不断增大，F 值始终未能超过前面的最优值. 故在神经网络训练时，选择 200 作为参数词向量维度 m 的值.

4.4　对比实验

实验中的对比模型如下.

(1) Trigram：使用 unigrams、bigrams 以及 trigrams 作为文本特征，分类器采用 SVM.

(2) TextFeatures：人工设计文本情感特征，输入到分类器 SVM 中[27].

(3) SSWE：通过 word2vec 学习特定的情感词向量，经过池化层的最大化平均化后输入到分类器 SVM 中[28].

(4) TextCNN-word：构建基于词语水平的 CNN 模型用于情感分类[8].

(5) TextCNN-char：构建基于字符水平的 CNN 模型用于情感分类[25].

(6) LSTM：不构建层次表示模型，直接将词向量输入到 LSTM 网络中进行情感分类[25].

(7) HLSTM：本文构建的去除注意力机制以

及去除 Chunk-Max Pooling 的层次表示模型.

(8) HLSTM-CMP:本文构建的去除注意力机制的层次表示模型.

(9) TSA:将文本与主题词向量结合输入到 LSTM 网络,使用基于上下文向量的注意力机制[29].

(10) HAN:分层注意力网络,分别基于 Attention 机制和 GRU 构建句子水平和文档水平的层次特征向量表示[18].

(11) TWE-ANN:本文提出的情感分析模型.

4.4.1 模型训练 基于以上介绍的 11 种方法来针对训练集进行模型的训练,并在训练过程中将模型在验证集上的性能表现 F 值进行对比,具体情况如下所示.

使用 TextCNN-word,TextCNN-char 与 LSTM 等三个没有应用层次表示特征的模型与基于传统方法的模型(Trigram、TextFeatures、SSWE)针对 Yelp2015 数据集的验证集进行实验所得到的对比结果如图 6 所示. 可以明显发现,神经网络模型较传统模型在前半段时间性能表现优越,但随着数据量的增大,传统模型的性能与神经网络模型渐渐接近,说明在大规模文本分类方面基于浅层神经网络的模型并没有什么优势.

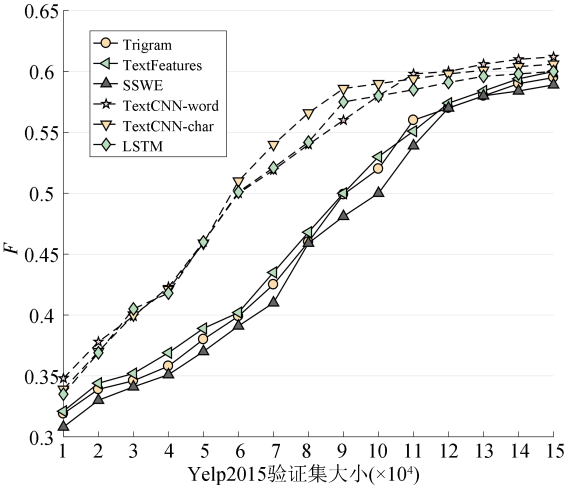

图 6 传统模型与深度网络模型性能对比

Fig. 6 Performance comparison between traditional model and deep network model

为了探索针对文档水平的情感分类层次表示模型的表现,本文将应用层次表示网络的模型 HLSTM 与没有应用层次表示网络的两个模型 TextCNN-word、LSTM 进行对比,针对 Amazon 验证集的表现如图 7 所示. 可以明显发现,基于层次表示的神经网络模型 HLSTM 在 Amazon 数据集上的情感分类效果明显优越于其他两个模型,说明在大规模文本分类方面层次表示模型对于模型性能改善起到很大的促进作用.

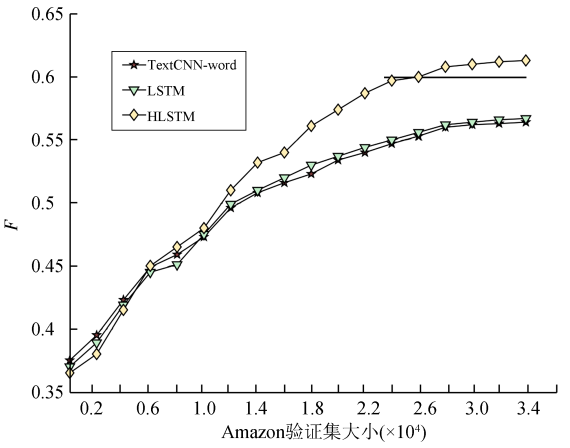

图 7 基于层次表示网络的模型性能对比

Fig. 7 Model performance comparison based on hierarchical representation network

使用 HLSTM、HLSTM-CMP 以及 TWE-ANN 三个模型针对 IMDB 的验证集进行性能对比,结果如图 8 所示. 将 HLSTM-CMP 与 HLSTM 模型进行对比,可发现加入了时序特征的模型性能表现更好,说明时序数据中特征的位置信息以及特征强度对模型的分类性能改善可以起到一定的促进作用. 本文提出的 TWE-ANN 模型进一步结合注意力机制与层次表示结构,与 HLSTM-CMP 进行对比,分类结果表明,基于主题增强的注意力机制的有效性,且可以促进模型分类性能的改善.

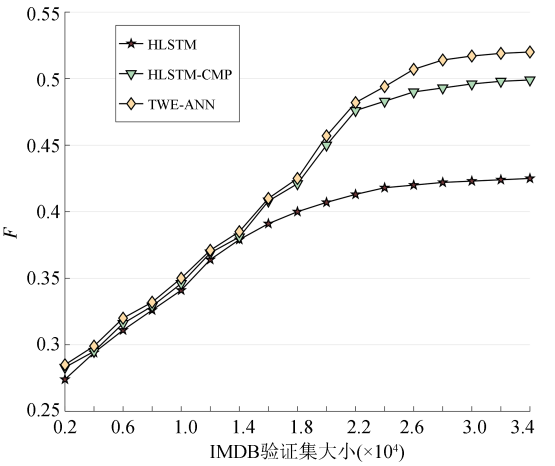

图 8 TWE-ANN 模型性能对比

Fig. 8 TWE-ANN model performance comparison

4.4.2 模型评估 基于以上介绍的 11 种训练好的模型,对三个测试集进行评估,表现效果如表 4.

表 4　实验结果与对比

Tab. 4　Experimental results and comparison

	Yelp2015		IMDB		Amazon	
	F	RMSE	*F*	RMSE	*F*	RMSE
Trigram	0.585	0.812	0.411	1.679	0.558	0.904
TextFeatures	0591	0.810	0.415	1.689	0.560	0.899
SSWE	0.579	0.855	0.367	1.891	0.551	0.911
TextCNN-word	0.610	0.798	0.397	1.755	0.554	0.807
TextCNN-char	0.600	0.811	0.387	1.763	0.568	0.812
LSTM	0.594	0.828	0.382	1.767	0.559	0.901
HLSTM	0.681	0.793	0.419	1.623	0.603	0.709
HLSTM-CMP	0.692	0.719	0.495	1.381	0.619	0.647
TSA	0.710	0.601	0.510	1.197	0.640	0.654
HAN	0.701	0.613	0.494	1.384	0.621	0.641
TWE-ANN	0.721	0.588	0.513	1.199	0.658	0.620

由表 4 可知，在大规模文本分类方面基于浅层神经网络的模型较传统机器学习模型并没有什么优势，与上一节介绍的一致. 例如 LSTM 在 Yelp2015、IMDB 以及 Amazon 数据集上的 *F* 值分别为 0.594、0.382、0.559，RMSE 分别为 0.828、1.767、0.901，而 TextFeatures 方法在三个数据集上的 *F* 值分别为 0.591、0.415、0.560，RMSE 分别为 0.810、1.689、0.899.

由表 4 可知，在大规模文本分类方面层次表示模型对于模型性能改善起到的作用很大，与上一节介绍的一致. 例如 HLSTM 较 LSTM 在三个数据集上的 *F* 值表现分别提升了 8.7%、3.7%、4.4%，在 RMSE 表现上分别提升了 3.5%、14.4%、19.2%. 对于时序特征以及主题注意力机制的性能改善也可通过表 4 中的相应模型对比发现.

为了进一步验证本文提出的模型性能优越表现，将 TSA、HAN 模型与 TWE-ANN 模型进行对比，可以发现：TWE-ANN 较其余两个模型在数据集上的最佳表现在 *F* 值上分别提升了 1.1%、0.3%、1.8%，在 RMSE 方面 TWE-ANN 在 Yelp2015，Amazon 数据集上分别提升了 1.3%、2.1%，而在 IMDB 数据集上有些许下降. 这可能是因为数据集样本的不平衡性以及数据不完全性导致. 综合意义上来说，这个结果也进一步验证了基于语料库的文档主题分布结合注意力机制对于层次表示情感分析模型的有效性以及性能改善性.

5　结　论

情感分类作为自然语言处理领域的重要研究方向之一，应用传统机器学习算法以及现下流行的神经网络算法总会在人工特征工程建立以及情感语义理解等方面存在信息缺失，故本文针对大规模文档语料基于 Attention 机制和双向 LSTM 神经网络结合主题模型提出了一种用于文档情感分析的深度模型. 首先基于 word2vec 进行词嵌入从而获取词向量表示，进而通过本文设计的 TWE-ANN 网络模型分别提取文档在词语级别和句子级别的深度情感特征，最后使用 softmax 进行情感分类. 实验结果表明在层次表示神经网络模型中加入基于主题增强的注意力机制能够有效地对大规模文本进行情感分类且在一定程度上促进了情感分类器性能的提升.

参考文献：

[1]　王仲远，程健鹏，王海勋，等. 短文本理解研究[J]. 计算机研究与发展，2016，53：262.

[2]　何炎祥，孙松涛，牛菲菲，等. 用于微博情感分析的一种情感语义增强的深度学习模型[J]. 计算机学报，2017，40：773.

[3]　Medhat W，Hassan A，Korashy H. Sentiment analysis algorithms and applications：a survey [J]. Ain Shams Engin J，2014，5：1093.

[4]　刘颖，贺聪，张清芳. 基于核相关分析算法的情感识别模型[J]. 吉林大学学报：理学版，2017，55：1539.

[5]　阳馨，蒋伟，刘晓玲. 基于多种特征池化的中文文本分类算法[J]. 四川大学学报：自然科学版，2017，54：287.

[6]　王永，陶娅芝，张勤. 中文网络评论中的产品特征情感倾向提取算法研究[J]. 重庆邮电大学学报：自然科学版，2017，29：75.

[7]　Li S S，Huang L，Wang R，*et al*. Sentence-level emotion classification with label and context dependence [C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics. [s. l.]：ACL，2015.

[8]　Kim Y. Convolutional neural networks for sentence classification [C/OL]//Proceedings of Conference on Empirical Methods in Natural Lan-guage Processing.（2014-09-03）[2018-05-25]. https://arxiv.org/abs/1408.5882.

[9]　Zhu X D，Parinaz S，Guo H Y. Long shortterm memory over recursive structures [C]//Proceedings of International Conference on International Conference on Machine Learning. New York：ACM Press，2015.

[10]　Phan X H，Nguyen C T，Le D T. A hidden topic-based frameworktoward building applications with short web documents [J]. IEEE Trans Knowl Data En，2011，23：961.

[11] Vo D T, Ock C Y. Learning to classify short text from scientific documents using topic models with various types of knowledge [C]. [S. l.]: Pergamon Press, Inc., 2015.

[12] Kalchbrenner N, Grefenstette E, Blunsom P. A convolutional neural network for modelling sentences [C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Boston, USA: MIT Press, 2015.

[13] Cheng X Q, Yan X H, Lan Y Y, *et al*. BTM: topic modeling over short texts [J]. IEEE Trans, 2014, 26: 2928.

[14] Ren Y F, Wang R M, Ji D H. A topic-enhanced word embedding for Twitter sentiment classification [J]. Inform Sciences, 2016, 369: 188.

[15] 刘全, 梁斌, 徐进, 等. 一种用于基于方面情感分析的深度分层网络模型[EB/OL]. (2017-01-17)[018-10-26]. http://kns.cnki.net/kcms/detail/11.1826.TP.20171129.2026.006.html.

[16] 支淑婷, 李晓戈, 王京博, 等. 基于多注意力长短时记忆的实体属性情感分析[EB/OL]. (2018-09-26)[2018-10-26]. http://kns.cnki.net/kcms/detail/51.1307.TP.20180926.1606.006.html.

[17] Yang Z C, Yang D Y, Chris D, *et al*. Hierarchical attenti-on networks for document classification [C]// Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics. California, USA: [s. n.], 2016.

[18] 胡朝举, 梁宁. 基于深层注意力的 LSTM 的特定主题情感分析[EB/OL]. (2018-0314)[2018-10-26]. http://kns.cnki.net/kcms/detail/51.1196.TP.20180314.1729.016.html.

[19] Zhou C, Sun C, Liu Z, *et al*. A C-LSTM Neural network for text classification [J]. Comput Sci, 2015: 1: 39.

[20] Klaus G, Rupesh Kumar S, Jan K, *et al*. LSTM: a search space odyssey [J]. IEEE Trans Neur Net Lear, 2017, 28: 2222.

[21] Chen Y, Xu L, Liu K, *et al*. Event extraction via dynamic multi-pooling convolutional neural networks [C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing. Boston, USA: MIT Press, 2015.

[22] 李平, 戴月明, 吴定会. 双通道卷积神经网络在文本情感分析中的应用[J]. 计算机应用, 2018, 38: 1542.

[23] Tang D Y, Qin B, Liu T. Document modeling with gated recurrent neural network for sentiment classification [C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Boston, USA: MIT Press, 2015.

[24] Diao Q M, Qiu M H, Wu Y, *et al*. Jointly modeling aspects, ratings and sentiments for movie recommendation(jmars) [C]//Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. New York: ACM Press, 2014.

[25] Li Y, Wang X T, Xu P J. Chinese text classification model based on deep learning [J]. Future Internet, 2018, 10: 11.

[26] Manning C D, Surdeanu M, Bauer J, *et al*. The stanford corenlp natural language processing toolkit [C]//Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Boston, USA: MIT Press, 2014.

[27] Kiritchenko S, Zhu X, Mohammad S M. Sentime-nt analysis of short informal texts [J]. J Artif Intell Res, 2014, 14: 723.

[28] Tang D Y, Wei F R, Yang N, *et al*. Learning sentiment-specific word embedding for twitter sentiment classification [C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Boston, USA: MIT Press, 2014.

[29] Baziotis C, Pelekis N, Doulkeridis C. DataStories at Semeval-2017 task 4: deep LSTM with attention for message-level and topic-based sentiment analysis [C]//Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). New York: ACM Press, 2017.

引用本文格式:

中 文: 刘广峰, 黄贤英, 刘小洋, 等. 基于主题注意力层次记忆网络的文档情感建模 [J]. 四川大学学报: 自然科学版, 2019, 56: 833.

英 文: Liu G F, Huang X Y, Liu X Y, *et al*. Document sentiment modeling based on topic attention hierarchy memory network [J]. J SichuanUniv: Nat Sci Ed, 2019, 56: 833.