软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2021,32(9):2783−2800 [doi: 10.13328/j.cnki.jos.005992]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于多通道特征和自注意力的情感分类方法*

李卫疆, 漆 芳, 余正涛

(昆明理工大学 信息工程与自动化学院,云南 昆明 650500)

通讯作者: 李卫疆, E-mail: hrbrichard@126.com

摘 要: 针对情感分析任务中没有充分利用现有的语言知识和情感资源,以及在序列模型中存在的问题:模型会将输入文本序列解码为某一个特定的长度向量,如果向量的长度设定过短,会造成输入文本信息丢失.提出了一种基于多通道特征和自注意力的双向 LSTM 情感分类方法(MFSA-BiLSTM),该模型对情感分析任务中现有的语言知识和情感资源进行建模,形成不同的特征通道,并使用自注意力重点关注加强这些情感信息.MFSA-BiLSTM 可以充分挖掘句子中的情感目标词和情感极性词之间的关系,且不依赖人工整理的情感词典.另外,在 MFSA-BiLSTM 模型的基础上,针对文档级文本分类任务提出了 MFSA-BiLSTM-D 模型.该模型先训练得到文档的所有的句子表达,再得到整个文档表示.最后,对 5 个基线数据集进行了实验验证.结果表明:在大多数情况下,MFSA-BiLSTM 和 MFSA-BiLSTM-D 这两个模型在分类精度上优于其他先进的文本分类方法.

关键词: 情感分类;多通道特征;自注意力;深度学习;双向 LSTM

中图法分类号: TP391

中文引用格式: 李卫疆,漆芳,余正涛.基于多通道特征和自注意力的情感分类方法.软件学报,2021,32(9):2783−2800. http://www.jos.org.cn/1000-9825/5992.htm

英文引用格式: Li WJ, Qi F, Yu ZT. Sentiment classification method based on multi-channel features and self-attention. Ruan Jian Xue Bao/Journal of Software, 2021,32(9):2783−2800 (in Chinese). http://www.jos.org.cn/1000-9825/5992.htm

Sentiment Classification Method Based on Multi-channel Features and Self-attention

LI Wei-Jiang, QI Fang, YU Zheng-Tao

(Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China)

Abstract: The purpose of this study is for the problem that the existing language knowledge and emotion resources are not fully utilized in the emotion analysis tasks, as well as the problems in the sequence model: the model will decode the input text sequence into a specific length vector, if the length of the vector is set too short, the information of input text will be lost. A bidirectional LSTM sentiment classification method is proposed based on multi-channel features and self-attention (MFSA-BiLSTM). This method models the existing linguistic knowledge and sentiment resources in sentiment analysis tasks to form different feature channels, and uses self-attention mechanism to focus on sentiment information. MFSA-BiLSTM model can fully explore the relationship between sentiment target words and sentiment polar words in a sentence, and does not rely on a manually compiled sentiment lexicon. In addition, this study proposes the MFSA- BiLSTM-D model based on the MFSA-BiLSTM model for document-level text classification tasks. The model first obtains all sentence expressions of the document through training, and then gets the entire document representation. Finally, experimental verifications are conducted on five sentiment classification datasets. The results show that MFSA-BiLSTM and MFSA-BiLSTM-D are superior to other state-of-the-art text classification methods in terms of classification accuracy in most cases.

Key words: sentiment classification; multi-channel features; self-attention; deep learning; bidirectional LSTM

* 基金项目: 国家自然科学基金(62066022); 国家重点研发计划(2018YFC0830105)

Foundation item: National Natural Science Foundation of China (62066022); National Key Research and Development Program of China (2018YFC 0830105)

收稿时间: 2019-06-24; 修改时间: 2019-10-31; 采用时间: 2019-12-11

随着深度学习技术的发展,基于神经网络的方法成为主流,被广泛地应用于自然语言处理(NLP)领域中.与传统的机器学习方法相比,深度学习在情感分析上表现得更为优秀,其不需要建立情感词典.深度学习能够自动捕捉从数据本身到高层更为复杂的语义映射,在性能上体现出比以往方法更好的效果.递归自动编码器[1,2]、卷积神经网络(CNN)[3−5]和长短期记忆网络(LSTM)[6,7]是目前在情感分析任务中常见的深度学习模型.

虽然这些神经网络模型在情感分类任务中取得了巨大的成功,但依然存在着一些缺陷:

首先,忽略了情感分析任务中现有的语言知识和情感资源,不能充分地利用这些情感特征信息;其次,语言知识(如情感词汇、否定词和程度副词等),在神经网络模型中未被充分使用.Chen 等人[8]提出了一种结合情感词典和卷积神经网络的情感分类方法(WFCNN),主要是利用情感词典中的词条对文本中的词语进行抽象表示,再使用卷积神经网络提取抽象词语的序列特征.该方法中的情感特征依赖于人工整理的情感词典,使用的特征单一,难以正确的表达每个词在句子中的重要程度,无法充分利用情感分析任务中语言知识和情感特征信息;并且该方法使用的 CNN 滤波器的词容量有限的,不能捕捉到远距离依赖,无法获得句子中非相邻词之间的语义关系.LSTM 可以通过对句子的顺序建模来解决这个限制.Qian 等人[9]提出了句级标注训练的 LSTM 模型,对情感词汇、否定词和程度副词等现有的语言规则进行建模,能够有效地利用语言学规则,实验也取得了较好的结果.但是,该模型需要大量的人力来建立强度正则化器.

另外,在深度学习中,很多的 NLP 任务都可以看作是一个序列建模任务(sequence modeling).而序列模型存在一个问题:无论输入的文本序列的长度为多少,最终都会将这个文本序列解码成为某一个特定的长度向量.如果设定的向量长度过短,那么会造成输入文本信息丢失,最后会导致文本误判.Pei 等人[10]针对这个问题提出了一种将词性注意力机制和 LSTM 相结合的网络模型,利用注意力矩阵计算出给定词句的注意力特征.实验结果表示:在一定的维度内,该模型能够取得较好的情感分类效果;但是,当文本映射的维度超过了阈值,分类的准确率会随着向量维度的提升而降低.Liu 等人[11]提出了一种具有注意机制和卷积层的双向 LSTM 文本分类模型,用来解决文本的任意序列长度问题,以及文本数据的稀疏问题.

针对以上问题,本文提出了一种基于多通道特征和自注意力的双向 LSTM 情感分类方法(MFSA-BiLSTM),模型由两部分组成:多通道特征和自注意力机制(self-attention).首先,本文对情感分析任务中现有的语言知识和情感资源进行建模,将输入文本句子中的词向量与词性特征向量,位置特征向量和依存特征向量三者进行结合形成不同的特征通道向量作为 BiLSTM 输入,让模型从不同的角度去学习句子中的情感特征信息,挖掘句子中不同方面的隐藏信息.然后,将这 3 个特征通道向量与 3 个 BiLSTM 的输出向量进行结合,再利用自注意力模型来发现句子中的重要信息,并对这些重要信息进行重点关注加强.本文采用的自注意力是注意力的一种特殊情况.与传统的注意力机制不同的是,自注意机制能够减少对外部信息的依赖,无视词与词之间的距离,直接计算依赖关系,学习每个词对句子情感倾向的权重分布,重点关注以及加强句子中的情感特征,可以使模型学习到更多的隐藏特征信息.本文的主要贡献如下.

(1) 本文经过研究发现,对情感分类任务中特有的语言知识和情感资源进行建模可以增强分类效果.本文通过在序列 BiLSTM 模型上建立多个特征通道向量输入来解决这个问题;
(2) 提出了一种自注意力机制.将多特征向量和 BiLSTM 模型的隐藏输出层相结合,为不同词赋予不同的情感权重.能够有效地提高了情感极性词的重要程度,充分挖掘文本中的情感信息;
(3) 同时,在本文提出的 MFSA-BiLSTM 模型基础上,本文提出了用于文档级文本分类任务的 MFSA-BiLSTM-D 模型;
(4) 在句级和文档级据集上验证了本文提出 MFSA-BiLSTM 模型和 MFSA-BiLSTM-D 模型在情感分析任务中的有效性.

1 相关工作

1.1 用于情感分析的语言知识

在情感分析任务中,语义知识和情感资源,例如情感词汇、否定词语(不、从不)、程度词(非常、绝对地)等

等,能够在很大程度上提高分类效果.因此,很多研究者尝试从语言知识和情感资源中设计出更好的特征来提高情感分析的分类性能.Tang 等人[12]将生成具有情感特定词嵌入(SSWE)的特征拿来训练 SVM 的分类模型.Huang 等人[13]将情感表情符号与微博用户性格情绪特征纳入到图模型 LDA 中实现微博主题与情感的同步推导,并在 LDA 中加入了情感层与微博用户关系参数[14],利用微博用户关系与微博主题来学习微博的情感极性.Vo 等人[15]在情感词典中添加表情特征用来自动构建文本,对 Twitter 文本进行情感分析.另外,还有一些关于从社交数据以及多种语言[16]中自动构建情感词典的研究.Teng 等人[17]提出了一种基于简单加权和上下文敏感词典的方法,使用 RNN 来学习情感强度,强化和否定词汇情感,从而构成句子的情感价值.将方面信息、否定词、短语情感强度、解析树及其组合应用到模型中以改进其性能.

但是众所周知,标准 RNN 会在其梯度下产生爆炸和消失状态.长短期记忆网络(LSTM)[6,7]是一种以长短期记忆单元为隐藏单元的 RNN 结构,能够有效地解决梯度消失和梯度爆炸问题.此外,LSTM 还考虑了词序列之间的顺序依赖关系,可以捕捉远距离的依赖,也可以捕获近距离的依赖.Tai 等人[18]提出一种将记忆细胞和门引入树形结构的神经网络模型 Tree-LSTM.Qian 等人[9]提出了语言规则化的 LSTM 模型(LR-Bi-LSTM),其中,情感词汇、否定词和强度词都被认为是句级情感分析的一个模型.Zhang 等人[19]提出一种基于批评学习和规则优化的卷积神经网络的情感分析,由基于特征的预测器、基于规则的预测器和批评学习网络这 3 个关键部分组成.其中,对于负极性规则和句子结构规则,模型需要人工去整理一个额外的情感词典(否定词和转折词).

与文献[9,17]相同的是,本文提出的 MFSA-BiLSTM 模型同样是对情感词汇,否定词和强度词等语言知识进行了建模.不同的是:MFSA-BiLSTM 模型对这些语言知识进行建模,形成不同的特征通道,让 BiLSTM 从不同的角度去学习句子中的特征信息;并且不需要大量的人工来建立强度正则化器[9]和整理一个额外的情感词典(否定词和转折词)[19],也不需要依赖解析树结构[17]以及昂贵的短语级注释的模型[18].

1.2 用于情感分类的注意力

目前,注意力机制已经成为一种选择重要信息以获取优异结果的有效方法.注意力机制最早是在计算机视觉领域提出来的,目的是模仿人类的注意力机制,给图像不同的局部赋予不同的权重.

Bahdanau 等人[20]在机器翻译任务上使用了注意力机制,是第一个将注意力机制应用到了 NLP 领域.Ma 等人[21]提出了一种基于隐藏状态的注意机制模型,该模型从上下文和方面交互式地学习注意力.Wang 等人[22]提出了基于注意的 LSTM 用于方面层面的情感分类与文献[23]中提出的基于内容注意的方面情感分类模型,关键思想都是向注意力机制添加方面信息.Liang 等人[24]提出一种基于多通道注意力卷积神经网络模型,用于特定目标情感分析.Guan 等人[25]使用的注意力机制直接从词向量的基础上学习每个词对句子情感倾向的权重分步,能够学习到增强情感分类效果的词语.Zhou 等人[26]提出的一种基于注意力的 LSTM 网络和 Vaswani 等人[27]提出的自注意力和多头注意力模型,都是用来解决跨语言的情感分类任务.Lin 等人[28]使用自注意力机制学习 LSTM 网络中句子的词嵌入,在情感分类任务上取得了较好的结果.Wang 等人[29]提出一种基于 RNN 的情绪分类胶囊,使用了注意力机制来构建胶囊表示.Liu 等人[11]提出了一种具有注意机制和卷积层的双向 LSTM 文本分类模型,使用注意力对 BiLSTM 隐层输出的信息进行不同的关注,解决文本的任意序列长度问题以及文本数据的稀疏问题.

与文献[11]中利用 LSTM 前一刻输出的隐含状态与当前时刻输入的隐藏状态进行对齐方式的注意力不同的是,MFSA-BiLSTM 模型使用的是直接对当前输入自适应加权的自注意力机制,无视词与词之间的距离,直接计算依赖关系,学习一个句子的内部结构.

2 基于多通道特征和自注意力的双向 LSTM 模型(MFSA-BiLSTM)

本文提出的模型总体架构如图 1 所示.形式上是以一个文本中词为单位,形成一个词序列:$\{x_1,x_2,\ldots,x_n\}$,每个词都通过已训练好的词向量映射成一个多维连续值的向量 w_i,$1\leqslant i\leqslant n$.再将句子序列中的词向量拼接,得到整个句子序列的词向量矩阵,表示为:$W^d=w_1\oplus w_2\oplus\ldots\oplus w_n$,维度为 d.模型不直接使用词向量 W^d 作为 BiLSTM 的输入,而是以词向量为基础分别与词性特征向量,位置值向量和依存句法向量进行组合形成不同的通道(见第 2.1

节),目的是为了让模型从不同角度去学习情感特征信息,充分地挖掘句子中的隐藏信息.

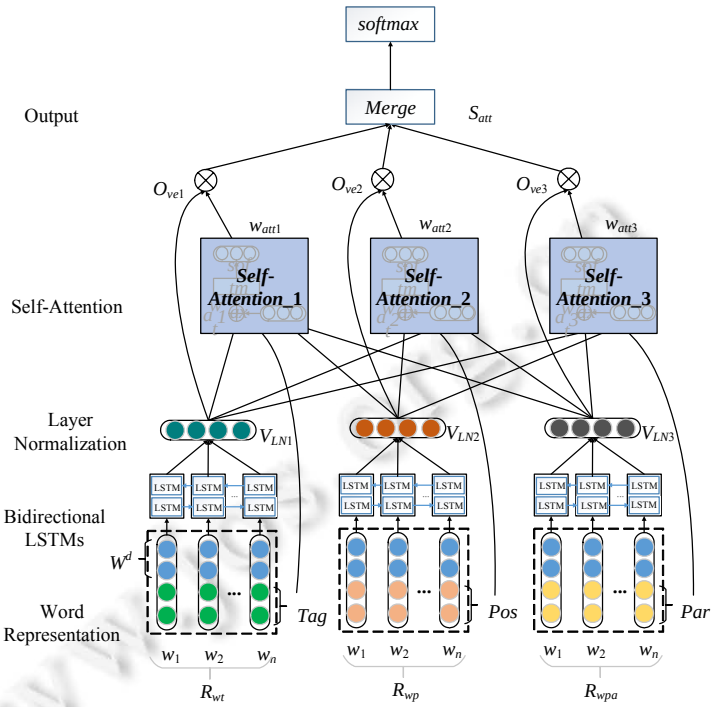

Fig.1 Architecture of the MFSA-BiLSTM

图 1 MFSA-BiLSTM 的体系结构

如图 1 所示,BiLSTM 提取了 3 个通道特征输入的特征信息,分别经过层归一化得到 V_{LN},再通过自注意力机制来学习一个加权矩阵 S_{att} 对原来的 V_{LN} 进行加权,为不同词赋予了不同的情感权重,从而进行情感分类.具体设计将在以下小节中介绍, MFSA-BiLSTM 的算法如下所示.

Algorithm 1. MFSA-BiLSTM 算法.

Input:使用后文公式(1)~公式(3),将词向量 W^d、词性向量 Tag^m、位置值向量 Pos^l 和依存句法向量 Par^p 构造成多通道特征输入;

Output:返回 p^k,其中,k 为任务.

for *iteration t* **do**

1: 使用后文公式(5)和公式(6),从多通道特征序列中获取前向后向上下文特征;

2: 使用后文公式(7)~公式(9)计算 BiLSTM 隐层中神经元的求和输入的均差和方差,得到隐层的输出 V_{LN};

3: 使用后文公式(11)~公式(13)来计算每个通道的词自注意力权重矩阵 w_{att};

4: 使用后文公式(14),对每个通道 BiLSTM 的隐层输出 V_{LN} 进行加权,即加权后的注意力特征向量为 O_{ve};

5: 将 3 个通道的注意力特征向量进行融合得到 S_{att},再利用 *softmax* 函数对其进行分类;

6: 最后使用后文损失函数公式(17)、Adadelta 方法来更新模型参数.

end

2.1 多通道特征

本文中的多通道特征由整个数据集中的词向量 W^d、词性特征向量 Tag^m、位置值向量 Pos^l 和依存句法向量 Par^p 构成.

- 词性特征向量.利用 HowNet 情感集合,对输入的句子中词语重新标注词性.通过词性标注,让模型去学

习对情感分类有重要影响的词语.其中,重点对特殊的情感词进行标注:程度副词(如非常、极其)、正面/负面评价词(如好、不好)、正面/负面情感词(喜欢、失望)和否定词(如不、从不).与词向量 W^d 操作一样,使 $t_i \in Tag^m$,其中,t_i 为第 i 个词性特征向量,m 是词性向量的维度;

- 位置值向量.在句子中,词与词之间的位置往往隐藏着重要信息,同一个词语出现在不同的位置,可能表达着不同的情感信息.将每个位置值映射成一个多维的连续值向量 $p_i \in Pos^l$,其中,p_i 为第 i 个位置特征向量,l 是位置特征向量的维度;
- 依存句法向量.依存句法分析是通过分析语言单位内成分之间的依存关系揭示其句法结构.通过对输入的句子进行句法分析,确定句子的句法结构和句子中词汇之间的依存关系,可以让模型在更大程度上学习情感分析任务中现有的语言知识,挖掘更多的隐藏情感信息.将每个句法特征映射成一个多维连续值向量 $parser_i \in Par^p$,其中,$parser_i$ 为句子 s 中第 i 个词的句法特征,p 是句法特征向量的维度.

接着,本文以词向量为基础,与词性特征向量,位置值向量和依存句法向量进行两两结合,形成 3 个通道作为网络模型的输入.让模型从不同角度去学习句子中不同方面的情感特征信息,挖掘句子中不同角度的隐藏信息.在实验中,本文使用一种简单行向量方向拼接操作:

$$R_{wt}=W^d \oplus Tag^m \tag{1}$$

$$R_{wp}=W^d \oplus Pos^l \tag{2}$$

$$R_{wpa}=W^d \oplus Par^p \tag{3}$$

2.2 长短期记忆网络和层归一化

长短期记忆网络(LSTM)[6,7]是对递归神经网络(RNN)的改进.在 LSTM 中,隐藏状态 h_t 和存储器单元 c_t 是之前的 h_{t-1} 和 c_{t-1} 和输入向量 W_t 的函数.每个位置(h_t)的隐藏状态只考虑前向,而不考虑后向,形式如下:

$$c_t,h_t=g^{LSTM}(c_{t-1},h_{t-1},W_t) \tag{4}$$

双向 LSTM[30]考虑前向和后向,学习两个方向的信息,能够更好地捕捉双向的语义依赖,如图 2 所示.

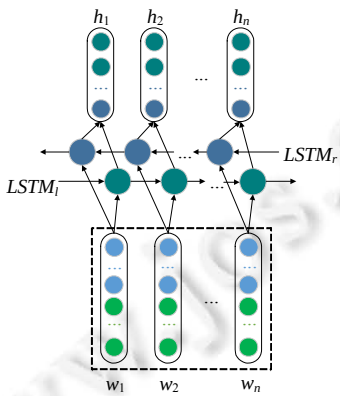

Fig.2 Bidirectional LSTM network structure

图 2 双向 LSTM 网络结构

双向 LSTM 是使用两个 LSTM 沿着序列的两个方向(前向和后向)扫描,并将两个 LSTM 的隐藏状态串联起来表示每个位置,前向和后向的 LSTM 分别表示为

$$\vec{c}_t,\vec{h}_t = g^{LSTM}(\vec{c}_{t-1},\vec{h}_{t-1},W_t) \tag{5}$$

$$\overleftarrow{c}_t,\overleftarrow{h}_t = g^{LSTM}(\overleftarrow{c}_{t+1},\overleftarrow{h}_{t+1},W_t) \tag{6}$$

其中,g^{LSTM} 与公式(4)中的相同,两个 LSTM 中的参数是共享的.整个句子的表示形式是 $[\vec{h}_n,\overleftarrow{h}_1]$,其中,$n$ 是句子词语的总个数.在位置 t 表示为 $h_t=\vec{h}_t \oplus \overleftarrow{h}_t$,这是前向 LSTM 和后向 LSTM 隐藏状态的级联.通过这种方式,可以同时考虑前向和后向上下文.

接下来,本文使用文献[31]提出的层归一化来计算隐藏层中神经元的求和输入的均差和方差,目的是稳定LSTM 网络中隐藏动态,防止模型过拟合.在层归一化中,本文对每个 BiLSTM 隐藏层 h_t 的每一个神经元赋予它们自己的自适应偏差和增益.层中的所有隐藏单元共享同样的归一化项μ和σ,形式如下:

$$h'_t = f\left[\frac{g}{\sigma^t}\odot(h_t-\mu_t)+b\right] \tag{7}$$

$$\mu_t = \frac{1}{H}\sum_{i=1}^{H} h_{t_i} \tag{8}$$

$$\sigma_t = \sqrt{\frac{1}{H}\sum_{i=1}^{H}(h_{t_i}-\mu_t)^2} \tag{9}$$

其中,H 为隐藏单元数量,⊙为两个向量之间的元素乘法,g 和 b 定义为与 h'_t 相同维度的偏差和增益参数.则BiLSTM 所有隐藏层状态的输出为公式(10),其中,V_{LN} 维度为 $n\times H$:

$$V_{LN} = (h'_1, h'_2, ..., h'_n) \tag{10}$$

2.3 自注意力机制

注意力机制最早是在图像处理领域提出来的,目的是为了在模型训练时,重点关注某些特征信息.常规的注意力机制做法是利用 LSTM 最后一个隐藏层的状态,或者是利用 LSTM 前一刻输出的隐层状态与当前输入的隐藏状态进行对齐.采用直接对当前输入自适应加权的自注意力,更合适用于情感分析任务中.

如表 1 所示,本文以词性特征为例对句子级 MR 数据集样例进行了分析.在样例中的情感词(如 impressively)能够体现出句子的情感倾向.为了加强这些情感词在分类时的作用,本文使用自注意力机制来学习一个句子的内部结构,重点加强句子中带有情感的特征信息.

Table 1 Analysis of key words in MR data samples

表 1 MR 数据样本关键词分析

MR 数据样本	关键词
An **ambitious**, **serious** film that manages to do **virtually** everything **wrong**; Sitting through it is something akin to an act of cinematic **penance**.	ambitious,serious,virtually, wrong,penance
Because of an **unnecessary** and **clumsy** last scene, 'swimfan' left me with a **very bad** feeling.	unnecessary,clumsy, very,bad
The emotion is **impressively true** for being so **hot-blooded**, and both leads are up to the task.	impressively,true, hot-blooded
The screenplay **sabotages** the movie's strengths at **almost** every juncture. All the characters are **stereotypes**, and their interaction is **numbingly predictable**.	sabotages,almost,stereotypes, numbingly,predictable

图 3 是 R_{wt} 通道的自注意力,其中,R_{wp} 通道的 V_{LN2} 和 R_{wpa} 通道的 V_{LN3} 作为额外辅助权值参与了 R_{wt} 通道的自注意力权重矩阵 w_{att1} 的计算:

$$\begin{cases} P_{VLN} = V_{LN1} \\ I_{tpp} = Tag^m \\ L_{nor} = L(V_{LN2}\oplus V_{LN3}) \end{cases} \tag{11}$$

$$a_{wt1} = P_{VLN}\oplus I_{tpp}\oplus L_{nor} \tag{12}$$

$$w_{att1} = softmax(L_3(\tanh(L_2(\tanh(L_1 a_{wt1}))))) \tag{13}$$

在上述公式中,P_{VLN},I_{tpp} 和 L_{nor} 为分别为自辅助矩阵、初始注意矩阵和额外辅助矩阵.L,L_1,L_2 和 L_3 分别是维度大小为 H,$3\times H+m+1$,$H+m$ 和 m 的权重,使用 *softmax* 进行归一化操作.然后,用自注意力权重 w_{att1} 对 BiLSTM 的隐藏状态 V_{LN1} 进行加权,即加权后的注意力特征向量 O_{ve1}:

$$O_{ve1} = w_{att1}\otimes V_{LN1} \tag{14}$$

与计算 R_{wt} 通道的注意力特征向量一样,得到 R_{wp} 和 R_{wpa} 通道的注意力特征向量为 O_{ve2} 和 O_{ve3}.情感分析本质上是一个分类问题,所以在模型的最后,将 3 个通道的注意力特征向量进行融合得到 S_{att},再利用 *softmax* 函数对其进行分类.如下:

$$S_{att}=[O_{ve1},O_{ve2},O_{ve3}] \tag{15}$$

$$p=softmax(w_cS_{att}+b_c) \tag{16}$$

其中,w_c 为权重矩阵,b_c 为偏置.在模型训练的过程中,本文使用交叉熵作为损失函数,且在模型参数上面使用权重衰减来对参数进行正则化.损失函数表示如下:

$$loss=-\sum_{i=1}^{D}\sum_{k=1}^{C}y_i^k\log p_i^k+\lambda\|\theta\|^2 \tag{17}$$

其中,D 为训练数据集大小,C 为数据的标签数,p 为预测的情感类别,y 为实际类别,$\lambda\|\theta\|^2$ 为 L2 正则项,λ为 L2 正则化超参数,θ为模型中的参数集.本文中使用时序反向传播算法(back propagation)来对网络参数进行更新.

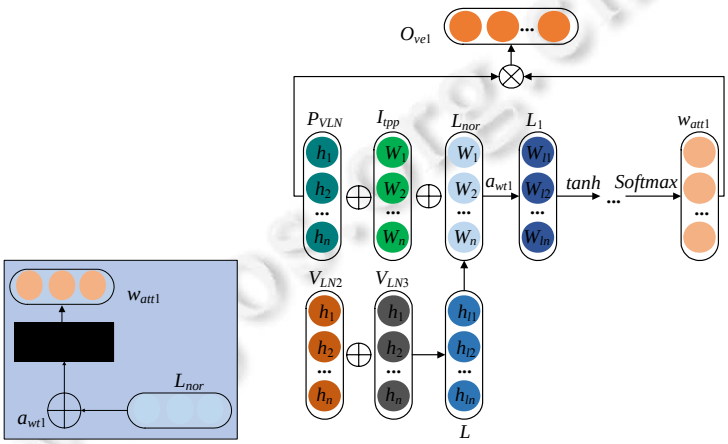

Fig.3 Self-Attention structure of R_{wt}

图 3 R_{wt} 通道的自注意力结构

2.4 MFSA-BiLSTM-D模型

在情感分类任务中,句子级文本的平均长度不超过 100(SL<100),见后文表 2.文本中的每个词可能具有一定的特征意义,会对分类结果产生影响.本文提出的 MFSA-BiLSTM 模型,充分学习了每个词语在句子中的语言特征信息,并且重点关注加强这些特征信息.因此,MFSA-BiLSTM 模型在句子级文本分类任务上效果显著(见后文表 4).然而,在平均长度超过 100(SL≥100)的文档级文本中,每个文本存在着多个句子,每个句子可能具有不同的情感倾向.所以,影响整个文档的分类效果是每个句子,而不是每个词语.

针对这一问题,Le 等人[32]提出了从句子和文档中学习分布式特征表示的无监督算法;Tang 等人[33]提出了将文档中每个用户和产品的文本偏好矩阵和表示向量引入 CNN 情感分类;Xu 等人[34]提出了一种缓存 LSTM 模型,用来捕获长文本中的整体语义信息;Chen 等人[35]在 LSTN 上使用了单词和句子级别的平均池层.

在本文中,若直接用 MFSA-BiLSTM 模型对文档级文本分类,会因为无法准确地获取文档中情感特征而导致分类效果不好(见后文表 5).因此,本文在 MFSA-BiLSTM 模型基础上,针对文档级文本分类任务提出了 MFSA-BiLSTM-D 模型(见图 4).与文献[32,35]一样,MFSA-BiLSTM-D 方法也是先训练得到句子表示,再得到文档表示.如图 4(左)所示,模型将文档 Doc.划分成为句子序列$[S_1,S_2,\ldots,S_m]$,其中,m 为句子个数;再将句子 S_i ($1\leqslant i\leqslant m$)划分为一系列单词$\{x_{i1},x_{i2},\ldots,x_{in}\}$,其中,$n$ 表示为 S_i 的长度.根据第 2.1 节对词进行特征向量化,形成 3 个通道;然后使用 MFSA-BiLSTM 模型学习文档中每个句子的词语情感,得到文档中每个句子表达向量 S_{attj} ($1\leqslant j\leqslant m$);接着,将 Doc.中的所有句子表达 $D_S=[S_{att1},S_{att2},\ldots,S_{attm}]$,送入如图 4(右)所示的模型进行训练.经过层归一化之后,计算句子自注意力权重矩阵 w_{satt}:

$$s_{wt}=V_{SLN}\oplus D_S \tag{18}$$

$$w_{satt}=softmax(L_2\tanh(L_1S_{wt})) \tag{19}$$

其中,V_{SLN} 为 BiLSTM 的隐藏输出;L_1 和 L_2 分别是维度大小为 H_S+m 和 m 的权重,H_S 为隐藏单元个数.最后得到

加权后的注意力特征向量 O_{sve}:

$$O_{sve}=w_{satt}\otimes V_{SLN} \tag{20}$$

最后,使用 *softmax* 函数对其进行分类.

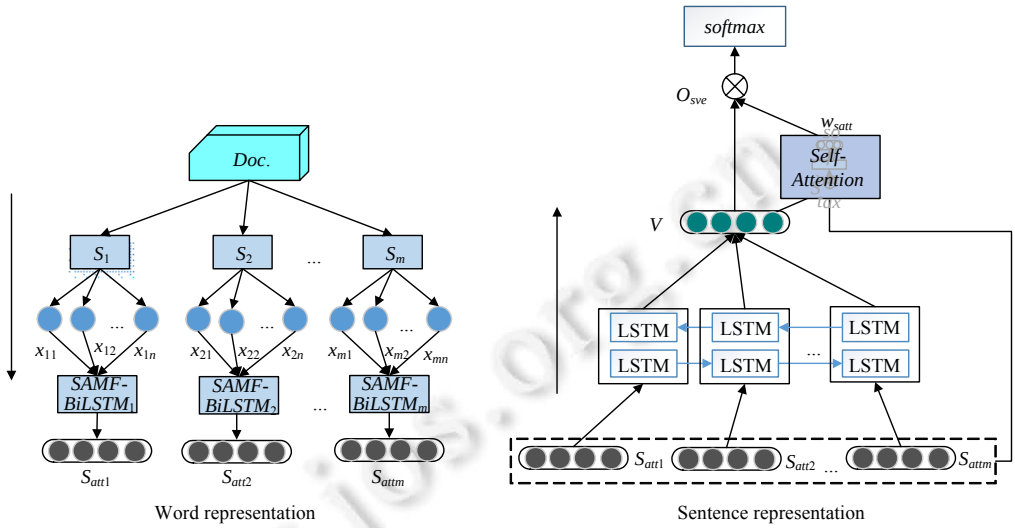

Fig.4 Architecture of MFSA-BILSTM-D

图 4 MFSA-BiLSTM-D 的体系结构

3 实验与分析

在本节中,本文在 5 个真实数据集下进行实验,展示了实验细节,评估了模型的性能并分析了结果.

3.1 数据集

(1) MR:MR 是一个二分类的电影评论数据集,包括 10 662 个样本,分别为 5 331 个正面和 5 331 个负面;

(2) SST-5:SST-5 是一个五分类数据集,是由斯坦福解析器在 11 855 个句子的解析树中解析的 227 376 个短语级细粒情感分类.本文在 SST-5 数据集上分别对句子级和基于短语级注释的句子级上进行训练,使用句子级中的测试数据进行测试;

(3) SST-2:在 SST-5 的数据集上进行整理(删除中性评论,非常积极和积极的评论被标记为积极,消极和非常负面的评论被标记为消极),得到二分类数据集 SST-2.本文在使用了短语级注释的 SST-2 数据集上进行训练,使用句子级中的测试数据进行测试;

(4) YELP3:来自 2013 年 Yelp 数据集挑战的评论数据集.每个评论的情绪极性是 1 星~5 星;

(5) IMDB:IMDB 是一个电影评论数据集,包括 84 919 个电影评论,范围从 1~10.

其中,MR,SST-5 和 SST-2 是句子级数据集(SL<100),YELP3 和 IMDB 是文档级数据集(SL≥100).表 2 显示了详细数据集的统计,其中,C 是目标类的数量,SL 是样本的平均长度,SD 表示文档中句子的平均数量,DS 是数据集的大小,WS 表示词汇量大小,$Test$ 是测试集的大小.

Table 2 Datasets for sentiment classification

表 2 情感分类的数据集

数据集	C	SL	SD	DS	WS	$Test$
MR	2	20	–	10 062	18 765	1 066
SST-5	5	18	–	11 855	17 836	2 210
SST-2	2	19	–	9 613	16 185	1 821
YELP3	5	189	11	71 193	48 957	8 671
IMDB	10	395	16	76 538	105 373	9 112

3.2 数据预处理与超参数设置

本文使用 Stanford CoreNLP 工具对表 2 的 5 个实验数据集进行分词、词性标注和依存句法分析.本文采用 Pennington 等人[36]提出的 Glove 向量作为单词嵌入的初始设置,其中每个词向量为 300 维,词典大小为 1.9MB.本文对 5 个实验数据集中的未登录词,使用均匀分布 $U(-0.05,0.05)$来随机初始化.在整个实验中,词向量维度为 300,词性特征为 30,位置特征为 25,依存句法特征为 25.训练过程本文使用 AdaDelta 梯度下降算法.所有数据集的 dropout rate 均设为 0.5.本文选择在测试数据集上表现最佳的结果作为最终表现.模型在不同数据集上参数设置见表 3.

Table 3 Optimal hyper-parameter configuration for five datasets

表 3 5 个数据集的最佳超参数配置

参数	MR	SST-5	SST-2	YELP3		IMDB	
				W	*S*	*W*	*S*
Learning rate	0.1	0.1	0.1	0.1	0.01	0.1	0.01
Hidden layer units	128	128	128	128	100	128	100
Weight Decay	1e−3	1e−4	1e−5	1e−4	1e−3	1e−4	1e−3
Batch Size	16	64	64	25	32	28	128

3.3 模型对比分析

将本文提出的两个模型分别与基准方法进行了比较,以验证本文提出的方法的有效性.基准方法可以分为 3 组,如下所示.

1. 一般基本模型

- SVM[37]:支持向量机;
- CNN[3]:使用预训练过的词嵌入的卷积神经网络模型;
- RNN[1]:循环神经网络;
- RNTN[2]:基于张量特征函数的情感树库上语义组合的递归深度神经网络;
- LSTM/BiLSTM:长短期记忆网络和双向长短期记忆网络;
- SSWE+SVM[12]:首先生成特定于情感的词嵌入来组成文档表示,然后训练 SVM 分类器;
- Paragraph-Vec[32]:从句子和文档中学习分布式特征表示的无监督算法;

2. 句子级网络模型

- Tree-LSTM[18]:将记忆细胞和门引入树形结构的长期短期记忆神经网络模型;
- NCSL[17]:将句子的情感分数视为句子中先前得分的加权和,其中,权重由神经网络学习得到;
- LR-Bi-LSTM[9]:语言规则化的 LSTM;
- RNN-capsule[29]:基于 RNN 的情绪分类胶囊模型;
- Capsule-B[38]:基于 CNN 的句子分类胶囊模型;
- AC-BiLSTM[11]:具有注意机制和卷积层的双向 LSTM 文本分类模型;
- CL+CNN[19]:一种基于关键学习情绪分析的正则卷积神经网络优化应用模型;

3. 文档级网络模型

- RNTN+RNN:用 RNTN 表示每个句子,并将句子表示输入 RNN;然后对 RNN 的隐藏向量进行平均,得到用于情绪分类的文档表示;
- UPNN(CNN)[33]:UPNN 将每个用户和产品的文本偏好矩阵和表示向量引入 CNN 情感分类,UPNN(CNN no UP)只使用 CNN,不考虑用户和产品信息;
- CIFG-LSTM/CIFG-BLSTM[39]:耦合输入忘记门 LSTM 和 BLSTM,分别表示为 CIFG-LSTM 和 CIFG-BLSTM.结合了 LSTM 的输入和遗忘门,与标准 LSTM 相比需要更少的参数;
- CLSTM[34]:缓存 LSTM 模型用来捕获长文本中的整体语义信息.这两种变体包括正则型和双向 B-CLSTM;

- NSC[35]:使用单词和句子级别的平均池层.NSC+LA 使用本地上下文捕获语义信息作为注意机制.

对比表 4 中句子级文本(MR,SST-5 和 SST-2)的实验结果.前 14 种方法的结果从文献[9,11,18,19]中引用.从表 4 中可以看出,MFSA-BiLSTM 在大多数基准数据集上取得了比其他方法更好的结果.在上述 14 种方法中,本文提出的方法优于除 MR 之外的所有数据集的其他基线.SST-5 和 SST-2 数据集上的 MFSA-BiLSTM 结果分别为 49.7%,51.8%和 89.7%.观察到:与 3 种基于 CNN 的方法(CNN,Capsule-B 和 CL+CNN)相比,MFSA-BiLSTM 在两个数据集上给出了更好的结果,说明本文使用的基于 LSTM 的方法比基于 CNN 的方法更适合此任务;同时,与两种都对语言知识进行建模的 LR-Bi-LSTM 方法和 NCSL 方法相比,MFSA-BiLSTM 方法的分类效果要更好,表明了本文提出对现有语言知识进行建模,生成不同的特征通道,让模型从不同角度的去学习句子中的情感特征信息的方法的有效性.与使用了注意力机制的 AC-BiLSTM 方法相比,本文使用的自注意力可以获得更好的性能.与依赖短语级注释的 Tree-LSTM 方法相比(当仅使用句子级进行训练时,其性能会下降 2.9%),MFSA-BiLSTM 方法不依赖于解析树,在使用了短语级注释和没有使用短语级注释的 SST-5 上的分类效果相差不大.另外,CL+CNN 方法在二分类 MR 数据集上是唯一一个达到 84.3%的方法.但是,本文提出的方法与 CL+CNN 的结果没有显著差异.同时,从表 4 还可以看出,基于深度学习方法的性能优于传统的机器学习方法.

Table 4 Experimental results of sentence-level sentiment classification accuracy

表 4 句子级情感分类准确性的实验结果

模型	MR	SST-5		SST-2
			+ phrase	+ phrase
SVM	–	–	40.7	79.4
Paragraph-Vec	–	–	48.7	87.8
CNN	81.5	46.9	48.0	87.2
RNN	77.7	43.2	44.8	82.4
RNTN	75.9	43.4	45.7	85.4
LSTM	78.3	45.6	46.4	84.9
BiLSTM	79.8	46.5	49.1	87.5
Tree-LSTM	80.7	48.1	51.0	88.0
NCSL	82.9	47.1	51.1	–
LR-Bi-LSTM	82.1	48.6	50.6	88.7
RNN-capsule	83.8	49.3	–	89.1
Capsule-B	82.1	48.6	–	88.7
AC-BiLSTM	83.2	48.9	–	88.3
CL+CNN	**84.3**	–	51.2	89.5
MFSA-BiLSTM	83.3	**49.7**	**51.8**	**89.7**

注:其中,实验结果通过分类准确度进行评估.省略%,"–"表示没有相关文献,该方法不使用该数据集.最佳结果以粗体显示

对比表 5 中文档级文本(YELP3 和 IMDB)的实验结果.前 13 种方法的结果从文献[33–35]中引用.

Table 5 Experimental results of document-level sentiment classification accuracy

表 5 文档级情感分类准确性的实验结果

模型	YELP3	IMDB
AvgWordvec+SVM	52.6	30.4
SSWE+SVM	54.9	31.2
Paragraph-Vec	55.4	34.1
RNTN+RNN	57.4	40.1
UPNN(CNN and no UP)	57.7	40.5
UPNN(CNN)	59.6	43.5
LSTM	53.9	37.8
BiLSTM	58.4	43.3
CIFG-LSTM	57.3	39.1
CIFG-BLSTM	59.2	44.5
CLSTM	59.4	42.1
B-CLSTM	59.8	46.2
NSC	62.7	44.3
NSC+LA	63.1	48.7
MFSA-BiLSTM	59.5	45.6
MFSA-LSTM-D	62.4	45.7
MFSA-BiLSTM-D	**63.8**	**48.9**

从表 5 中可以看出,本文提出的 MFSA-BiLSTM-D 方法比两个数据集上的其他基线获得了更好的结果(63.8%和 48.9%).与同样先是训练得到句子表示再得到文档表示的 RNTN+RNN 方法、Paragraph-Vec 方法、NSC 方法和 NSC+LA 相比,MFSA-BiLSTM-D 方法取得了更好的分类效果.这表明了本文提出的方法的有效性.同时,与改变 LSTM 模型的内部存储的 CIFG-LSTM,CIFG-BLSTM,CLSTM 和 B-CLSTM 相比,MFSA-BiLSTM-D 方法具有可行性.另外,从表 5 中可以看出:对于文档文本数据集(YELP3 和 IMDB),本文提出的 MFSA- BiLSTM-D 方法结果更优于句级 MFSA-BiLSTM 方法.这表明了 MFSA-BiLSTM-D 的方法比 MFSA-BiLSTM 方法更适合此任务.MFSA-BiLSTM-D 能够很好地捕获文档级文本的情感倾向.

3.4 自注意力机制和语言特征的影响

MFSA-BiLSTM 包括两个部分,即自注意力机制和多通道语言特征.对于 MFSA-BiLSTM,应该证明所有成分均可用于最终结果.在本节中,我们将进行一组实验来评估自注意力和多通道语言特征分别对 MFSA-BiLSTM 和 MFSA-BiLSTM-D 两个模型性能的影响.由于 MFSA-BiLSTM 不依赖于解析树,在使用了短语级注释过的和没有使用短语级注释过的 SST-5 上的分类效果相差不大.因此,为了统一分析,在后面所有实验中,对于 SST-5 数据集,本文只使用了短语注释过的 SST-5 数据集.

(1) 自注意力的影响

本文提出的词自注意力权重是由初始注意矩阵 I_{tpp}、自辅助矩阵 P_{VLN} 和额外辅助矩阵 L_{nor} 这 3 个部分构成(见图 3).为了揭示自注意力对模型的影响,在实验过程中保留了模型的语言特征部分.本文在 5 个数据集上对 MFSA-BiLSTM 和 MFSA-BiLSTM-D 两个模型分别进行自注意力权重调节实验.观察到结果见表 6 和表 7.

Table 6 Accuracy for MFSA-BiLSTM with different self-attention weights

表 6 不同自注意权重下 MFSA-BiLSTM 的精度

	MR	SST-5	SST-2
MF-BiLSTM	81.9	49.5	88.0
MFSA-BiLSTM(no I_{tpp})	82.3	50.8	88.4
MFSA-BiLSTM(no P_{VLN})	82.5	51.3	88.9
MFSA-BiLSTM(no L_{nor})	83.0	51.5	89.2
MFSA-LSTM(all)	82.2	51.1	88.6
MFSA-BiLSTM(our model)	**83.3**	**51.8**	**89.7**

Table 7 Accuracy for MFSA-BiLSTM-D with different self-attention weights

表 7 不同自注意权重下 MFSA-BiLSTM-D 的精度

	YELP3	IMDB
MF-BiLSTM-D	59.6	45.4
MFSA-BiLSTM-D(no I_{tpp})	63.0	47.6
MFSA-BiLSTM-D(no P_{VLN})	62.8	46.9
MFSA-BiLSTM-D(no L_{nor})	63.2	48.1
MFSA-BiLSTM-D(no w_{satt})	63.6	48.8
MFSA-BiLSTM-D(our model)	**63.8**	**48.9**

从表 6 和表 7 可以看出,完全不使用词注意力机制的 MF-BiLSTM 和 MF-BiLSTM-D 分类效果明显不如使用了词自注意力机制的 MFSA-BiLSTM(no I_{tpp})和 MFSA-BiLSTM-D(no I_{tpp})模型.这意味着自注意力对我们的方法有一定的影响.通过调节自注意力的权重,可以观察到:计算自注意力权重的初始注意矩阵 I_{tpp}、自辅助矩阵 P_{VLN} 和额外辅助矩阵 L_{nor},对 MFSA-BiLSTM 和 MFSA-BiLSTM-D 的性能有很大的影响.另外,在使用了完整自注意力权重的情况下,MFSA-LSTM(all)的分类效果明显不如 MFSA-BiLSTM(our model).可见,BiLSTM 能够比 LSTM 更好地解决序列建模任务.同时,拥有完整自注意力权重的 MFSA-BiLSTM(our model)和 MFSA-BiLSTM-D(our model)可获得最佳结果.它证明了自注意力中所有成分对于 MFSA-BiLSTM 和 MFSA-BiLSTM-D 的最终结果都是有用的.

(2) 不同语言特征的影响

本文提出的多通道语言特征包括 R_{wp}(由词向量和位置值组成),R_{wpa}(由词向量和句法组成)和 R_{wt}(由词向量

和词性向量组成)(如图 1 所示).为了揭示语言特征对模型的影响,本文在 5 个数据集上对 MFSA-BiLSTM 和 MFSA-BiLSTM-D 这两个模型分别进行了语言特征调节实验.

从表 8 和表 9 可以看出:随着语言特征的添加,模型的复杂度越来越高,模型的性能起伏比较大,但是模型的总体性能随着语言特征的添加呈上升趋势.使用 3 个通道的 MFSA-BiLSTM 和 MFSA-BiLSTM-D 比只使用了词特征的模型的分析提升了 1.8%~4.4%,其中,R_{wt} 和 R_{wpa} 在性能提升方面起着关键性的作用.这证明了多通道语言特征可以进一步提高 MFSA-BiLSTM 和 MFSA-BiLSTM-D 的性能.

Table 8 Accuracy for MFSA-BiLSTM with different linguistic feature

表 8 语言特征下 MFSA-BiLSTM 的准确性

	特征通道			MR	SST-5	SST-2
	R_{wp}	R_{wpa}	R_{wt}			
SA-BiLSTM	√	×	×	79.1	49.7	87.8
	×	√	×	80.9	50.2	88.3
	×	×	√	82.1	50.8	88.7
	√	√	×	81.9	50.5	88.5
	√	×	√	83.0	51.0	88.8
	×	√	√	82.9	51.4	89.4
	√	√	√	**83.3**	**51.8**	**89.7**

Table 9 Accuracy for MFSA-BiLSTM-D with different linguistic feature

表 9 不同语言特征下 MFSA-BiLSTM-D 的准确性

	特征通道			YELP3	IMDB
	Rwp	*Rwp*	*Rwt*		
SA-BiLSTM-D	√	×	×	59.4	45.8
	×	√	×	60.3	46.3
	×	×	√	61.7	47.7
	√	√	×	62.9	47.4
	√	×	√	63.1	47.9
	×	√	√	63.5	48.4
	√	√	√	**63.8**	**48.9**

3.5 向量大小和不同词嵌入的影响

从语言特征调节实验中,得出了在词向量的基础上,词性特征与句法特征在分类效果上起着关键性作用.因此,在这一小节对词性特征、句法特征以及词向量进行了进一步分析.

在图 5 和图 6 中展示了具有不同维度词性特征和句法特征大小的 MFSA-BiLSTM 和 MFSA-BiLSTM-D 模型性能.本文使用以下集合中的向量大小{10,20,25,30,50,100,200}.从图 5 可以看出:当词性向量大小变化时,模型在 MR,SST-2,YELP3 和 IMDB 这 4 个数据集都呈现上升的趋势.当词性向量>30 时,模型在 MR 和 SST-2 数据集上出现了波动;并且随着维度的增加,分类准确率呈现下降趋势.在 YELP3 和 IMDB 数据集上,模型性能趋于稳定.如图 6 所示:当句法向量>25 时,模型性能趋于稳定.因此,选择适合的词性向量和句法向量维度大小可以获得更好的结果.

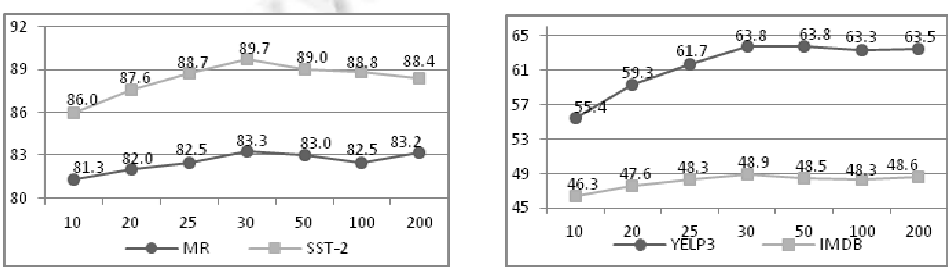

Fig.5 Influence of parts-of-speech features in different dimensions

图 5 词性特征在不同维度上的影响

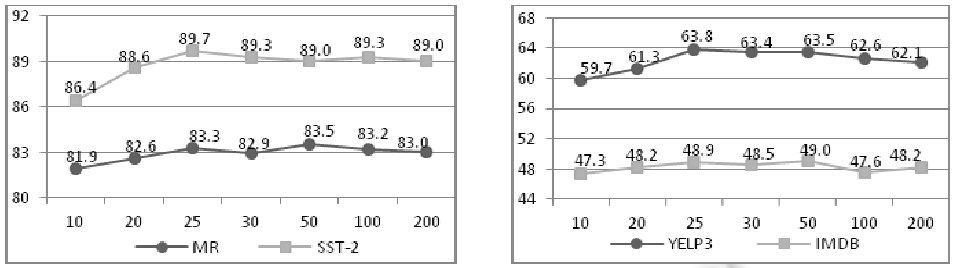

Fig.6 Influence of dependency parsing features in different dimensions

图 6 句法特征在不同维度上的影响

在图 7 中展示了 MFSA-BiLSTM 和 MFSA-BiLSTM-D 两个模型在不同维度下和不同初始词嵌入下的性能.本文使用以下集合中的向量大小{50,100,150,200,300},并设置预训练和随机两种初始词嵌入.注意,模型中所有单元的尺寸也会随之变化.从表 7 中可以看出:在所有数据集上,使用预训练的词嵌入向量的 MFSA- BiLSTM 和 MFSA-BiLSTM-D 比使用随机字嵌入向量的 MFSA-BiLSTM 和 MFSA-BiLSTM-D 效果更好.当向量大小变化时,使用预训练的词嵌入向量模型的性能都呈现稳定上升的趋势;而使用随机词嵌入向量的模型在向量>150 时,开始出现波动.与随机词嵌入向量相比,预训练词嵌入向量具有明显的优势.

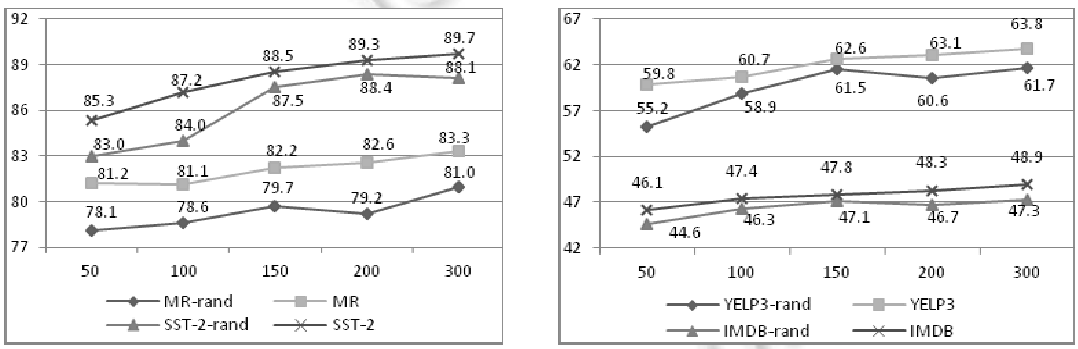

Fig.7 Influence of different word embedding and vector size

图 7 不同的词嵌入和向量大小的影响

3.6 不同文本长度的影响

在序列模型中,会将输入文本序列解码为某一个特定的长度向量,若向量的长度设定过短,可能会造成文本信息的丢失,导致文本理解出现偏差.针对这一问题.本小节在电影评论数据集(MR)进行了文本长度调节实验.

在实验中,根据电影评论数据集(MR)可视化(如图 8(左)所示).设定文本长度为 15~60,间隔为 5.在 LSTM,BiLSTM,MF-BiLSTM(无自注意力机制,有多通道特征)和本文提出的 MFSA-BiLSTM 等 4 个序列模型上进行了实验.实验结果如图 8(右)所示:当文本长度小于 35 时,LSTM,BiLSTM 和 MF-BiLSTM 这 3 个模型的分类性能急速下降;当文本长度大于 35 时,LSTM,BiLSTM 和 MF-BiLSTM 这 3 个模型的分类性能较平缓或呈缓慢上升趋势.本文提出的 MFSA-BiLSTM 模型的分类性能总体较稳定呈平缓趋势,当文本长度大于 50 时,MFSA-BiLSTM 模型的分类性能呈下降趋势.

因此,经实验分析可以看出:本文提出的 MFSA-BiLSTM 模型,在文本长度调节过程中的分类效果相差并不是很大.原因是 MFSA-BiLSTM 模型中的自注意力是由自辅助矩阵、初始注意矩阵和额外辅助矩阵这 3 部分组成,其中,初始注意矩阵能够在一定程度考虑到文本长度.但是,当文本长度超过一定阈值时,由于数据稀疏问题,MFSA-BiLSTM 模型的分类性能会受到影响.

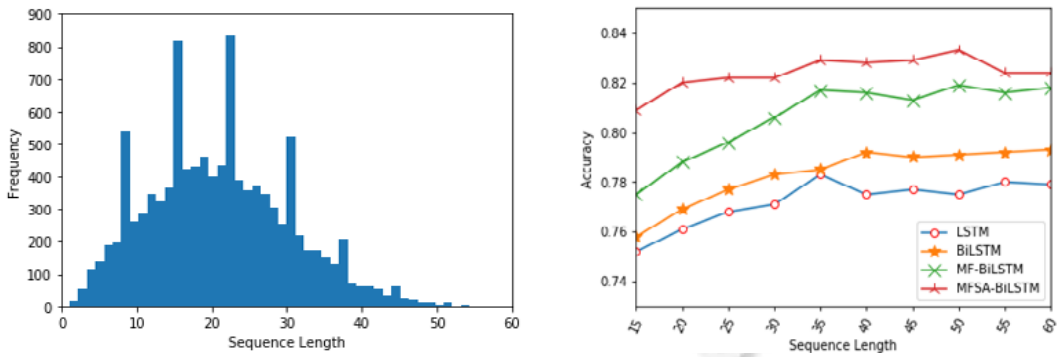

Fig.8 Movie Review(MR) dataset visualization (left) and accuracy of different text lengths (right)

图 8 电影评论数据集(MR)可视化(左)和不同文本长度下的精度(右)

4 案例分析与自注意力可视化

4.1 案例分析

为了进一步分析本文提出的模型相对于 BiLSTM(无自注意力,无多通道特征),MF-BiLSTM(无自注意力机制,有多通道特征),WFCNN(使用了情感序列特征的 CNN)以及 LR-Bi-LSTM(使用了语言特征的 LSTM)等模型的优势,本文使用经过训练的 MFSA-BiLSTM,BiLSTM,MF-BiLSTM,WFCNN 和 LR-Bi-LSTM 预测几个具体的样例来进行分析.由于 MFSA-BiLSTM-D 是在 MFSA-BiLSTM 上提出的,因此,在本节,本文只对 MFSA-BiLSTM 进行分析.

如表 10 样例分类结果所示.

Table 10 Analysis of typical sample cases

表 10 典型样例分析

ID	样例	目标	模型	判断结果
1	The last case (a different brand) we ordered from Amazon was so terrible we threw away the entire case. However, the Boscoli brand is good and we can enjoy a good dirty martini.	1	MFSA-BiLSTM LR-Bi-LSTM MF-BiLSTM BiLSTM WFCNN	√ × √ × ×
2	All of the elements are in place for a great film noir, but director george hickenlooper's approach to the material is too upbeat.	0	MFSA-BiLSTM LR-Bi-LSTM MF-BiLSTM BiLSTM WFCNN	√ × × × ×
3	After discovering the use of Samsung mobile phones, my Weibo is full of typos! Can't stand it! Be careful! Be careful!	0	MFSA-BiLSTM LR-Bi-LSTM MF-BiLSTM BiLSTM WFCNN	√ √ √ × ×

对于样例 3,情感词不是单独起作用的,而是通过词序列结合句子的上下语义表达出整个句子的情感.由于 WFCNN 提取的特征是局部相邻词之间的特征,因此出现误分类.BiLSTM 虽然具有强大的上下文语义捕捉能力,但是样例 3 具有大量的正负面情感词,由于对特殊的情感词并没有进行处理,从而出现了误分类.而 MFSA-BiLSTM,LR-Bi-LSTM 和 MF-BiLSTM 这 3 个模型充分利用了语言知识,不仅有强大的上下文语义捕捉能力,并能根据上下语义对文本中的情感词进行程度加强,因此能够正确分类.对于样例 1 和样例 2,这种带有“however”“but”转折词的文本,LR-Bi-LSTM 并没有分类成功.原因是 LR-Bi-LSTM 模型的调节器具有局限性,它没有考虑句子的依赖关系,而直接对整个文本的情感词进行强度调节.MF-BiLSTM,能够根据句子结构、词的位置和词性特征对一些带有转折词的文本进行正确的分类(样例 1),但当遇到分类情感特征不明显且带转折的文本时,会分

类错误(样例 2).而本文提出的 MFSA-BiLSTM 在 MF-BiLSTM 模型上增加了自注意力,通过自注意加权,加强文本中的情感,使情感特征信息特征更加突出.因此,本文提出的 MFSA-BiLSTM 模型可以分类成功.

4.2 自注意力可视化

本文在图 9 中可视化了 MR 数据的测试集中的两个案例,来解释 MFSA-BiLSTM 的多通道自注意力是如何工作的.颜色深度表示相应单词的重要程度.颜色越深,单词越重要.O_{ve1}、O_{ve2}、O_{ve3} 分别表示为文本经过 3 个通道自注意的得分向量.其中,图 9(a)的极性是正面,MFSA-BiLSTM 模型预测为正面;图 9(b)的极性是正面,MFSA-BiLSTM 模型预测为负面.

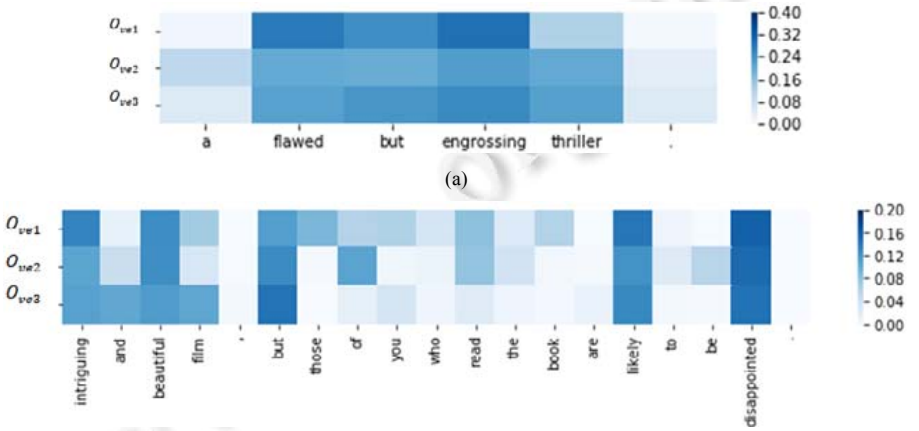

Fig.9 Three channel features self-attention visualization

图 9 3 个通道特征自注意可视化

如图 9 所示,图 9(a)是一个带有“but”子句的样例,样例的极性由“but”引导的句子决定.可以观察到,O_{ve1} 的注意力得分向量突出了“flawed”和“engrossing”两个情感比较明显的词.对于 O_{ve2} 的注意力得分向量,借助了位置信息以及词性和句法信息作为辅助,突出了“engrossing”,同时没有分散无关词的注意力.对于 O_{ve3} 的注意力得分,借助了句子中的句法以及词性和位置作为辅助,对“but”进行了转折加强,并影响到了“engrossing”,所以“engrossing”比“flawed”的颜色深一点.故 MFSA-BiLSTM 能够对样例进行正确预测.图 9(b)同样是一个带有“but”子句的样例.一般来说,在没指定目标词的情况下,样例的极性由“but”引导的句子决定.从整个样例来说,该样例的极性是负面的.但是在这个样例中,存在“film”和“book”两个目标词,以“film”为样例目标词,那么样例则判为正面.若以“book”为样例目标词,那么样例则判为负面.然而,该样例是属于 MR 数据集,MR 是一个电影评论数据集,因此,该样例要以“film”为目标词,判为正面.如图可见,MFSA-BiLSTM 并没有考虑以“film”目标词为预测中心,而是从句子结构出发,重点关注了“but”后面的子句,进行了错误的判断.

4.3 错误分析

为了更好地理解本文提出模型的局限性,本文对 MFSA-BiLSTM 模型所产生的误差进行了分析.具体来说,本文从 MR 电影评论数据集的测试集中随机选择了 50 个被 MFSA-BiLSTM 错误预测的实例,揭示了分类错误的几个原因.可以将其分为以下两种.

- 第 1 种,MFSA-BiLSTM 无法对存在多个目标词的文本进行正确的预测.例如对于一个句子“intriguing and beautiful film, but those of you who read the book are likely to be disappointed.”,会因为无法确定目标词是“film”还是“book”,本文提出的模型会直接根据句子的结构、位置以及词性,以“but”后面的“book”为目标词进行预测,从而出现误判;
- 第 2 种,当文本长短相差过大,会造成多通道特征稀疏,影响自注意力权重的分布,从而影响分类效果.

5 总结和未来工作

本文提出了一个具有自注意力机制和多通道特征的双向 LSTM 模型(MFSA-BiLSTM).该模型由自注意力机制和多通道特征两部分组成.先对情感分析任务中现有的语言知识和情感资源进行建模,生成不同的特征通道作为模型的输入,再利用 BiLSTM 来充分的获得这些有效的情感资源信息.最后使用自注意力机制对这些重要信息进行重点关注加强,提高分类精度.此外,本文在 MFSA-BiLSTM 模型上,针对文档级文本分类任务提出了 MFSA-BiLSTM-D 模型.该模型将文本中的句子进行分割,再分别使用 MFSA-BiLSTM 模型进行特征学习得到句子特征信息.在 5 个基准数据集上进行了实验,用来评估本文提出的方法的性能.实验结果表明:在大多数情况下,MFSA-BILSTM 和 MFSA-BILSTM-D 模型比一些最先进的基线方法分类更好.

未来的工作重点是注意力机制的研究和文档级文本特定目标分类任务的网络模型体系结构的设计.未来的工作主要包括以下几个部分:(1) 利用其他注意机制进一步完善本文提出的方法;(2) 针对文档级文本特定目标分类任务,设计了一种新的注意机制和网络模型;(3) 将本文的方法应用到实际应用中.

References:

[1] Socher R, Pennington J, Huang EH, Ng AY, Manning CD. Semi-supervised recursive autoencoders for predicting sentiment distributions. In: Proc. of the 2011 Conf. on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2011. 151−161.

[2] Socher R, Perelygin A, Wu J, Chuang J, Manning CD, Ng A, Potts C. Recursive deep models for semantic compositionality over a sentiment treebank. In: Proc. of the 2013 Conf. on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2013. 1631−1642.

[3] Kim Y. Convolutional neural networks for sentence classification. In: Proc. of the 2014 Conf. on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg: Association for Computational Linguistics, 2014. 1746−1751. [doi: 10.3115/v1/D14-1181]

[4] Kalchbrenner N, Grefenstette E, Blunsom P. A convolutional neural network for modelling sentences. In: Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2014. 655−665. [doi: 10.3115/v1/P14-1062]

[5] Lei T, Barzilay R, Jaakkola T. Molding CNNs for text: Non-linear, non-consecutive convolutions. In: Proc. of the 2015 Conf. on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2015. 1565−1575. [doi: 10.18653/v1/D15-1180]

[6] Zhu X, Sobihani P, Guo H. Long short-term memory over recursive structures. In: Proc. of Int'l Conf. on Machine Learning. 2015. 1604−1612.

[7] Tai KS, Socher R, Manning CD. Improved semantic representations from tree-structured long short-term memory networks. In: Proc. of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th Int'l Joint Conf. on Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2015. 1556−1566. [doi: 10.3115/v1/P15-1150]

[8] Chen Z, Xu RF, Gui L, Lu Q. Combining convolutional neural networks and word sentiment sequence features for Chinese text sentiment analysis. Journal of Chinese Information Processing, 2015,29(6):172−178 (in Chinese with English abstract). http://jcip.cipsc.org.cn/CN/Y2015/V29/I6/172 [doi: CNKI:SUN:MESS.0.2015-06-024]

[9] Qian Q, Huang M, Lei J, Zhu X. Linguistically regularized LSTM for sentiment classification. In: Proc. of the 55th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2017. 1679−1689. [doi: 10.18653/v1/P17-1154]

[10] Pei SW, Wang LL. Text sentiment analysis based on attention mechanism. Computer Engineering and Scienc, 2019,41(2): 344−353 (in Chinese with English abstract). [doi: CNKI:SUN:JSJK.0.2019-02-023]

[11] Liu G, Guo J. Bidirectional LSTM with attention mechanism and convolutional layer for text classification. Neurocomputing, 2019, 337:325−338. [doi: 10.1016/j.neucom.2019.01.078]

[12] Tang D, Wei F, Yang N, Zhou M, Liu T, Qin B. Learning sentiment-specific word embedding for twitter sentiment classification. In: Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2014. 1555−1565. [doi: 10.3115/v1/P14-1146]

[13] Huang FL, Feng S, Wang D, Yu G. Mining topic sentiment in microblogging based on multi-feature fusion. Chinese Journal of Computers, 2017,40(4):872−888 (in Chinese with English abstract). [doi: 10.11897/SP.J.1016.2017.00872]

[14] Huang FL, Yu G, Zhang JL, Li CX, Yuan CA, Lu JL. Mining topic sentiment in micro-blogging based on micro-blogger social relation. Ruan Jian Xue Bao/Journal of Software, 2017,28(3):694−707 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5157.htm [doi: 10.13328/j.cnki.jos.005157]

[15] Vo DT, Zhang Y. Don't count, predict! An automatic approach to learning sentiment lexicons for short text. In: Proc. of the 54th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2016. 219−224. [doi: 10.18653/v1/P16-2036]

[16] Chen Y, Skiena S. Building sentiment lexicons for all major languages. In: Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2014. 383−389. [doi: 10.3115/v1/P14-2063]

[17] Teng Z, Vo DT, Zhang Y. Context-sensitive lexicon features for neural sentiment analysis. In: Proc. of the 2016 Conf. on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2016. 1629−1638. [doi: 10.18653/v1/D16-1169]

[18] Tai KS, Socher R, Manning CD. Improved semantic representations from tree-structured long short-term memory networks. In: Proc. of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th Int'l Joint Conf. on Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2015. 1556−1566. [doi: 10.3115/v1/P15-1150]

[19] Zhang B, Xu X, Li X, Chen X, Ye Y, Wang Z. Sentiment analysis through critic learning for optimizing convolutional neural networks with rules. Neurocomputing, 2019,356:21−30. [doi: 10.1016/j.neucom.2019.04.038]

[20] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473, 2014.

[21] Ma D, Li S, Zhang X, Wang H. Interactive attention networks for aspect-level sentiment classification. In: Proc. of the 26th Int'l Joint Conf. on Artificial Intelligence. AAAI, 2017. 4068−4074.

[22] Wang Y, Huang ML, Zhao L, Zhu XY. Attention-based LSTM for aspect-level sentiment classification. In: Proc. of the 2016 Conf. on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2016. 606−615. [doi: 10.18653/v1/D16-1058]

[23] Liu Q, Zhang H, Zeng Y, Huang Z, Wu Z. Content attention model for aspect based sentiment analysis. In: Proc. of the 2018 Int'l Conf. on World Wide Web. Steering Committee, 2018. 1023−1032. [doi: 10.1145/3178876.3186001]

[24] Liang B, Liu Q, Xu J, Zhou Q, Zhang P. Aspect-based sentiment analysis based on multi-attention CNN. Journal of Computer Research and Development, 2017,54(8):1724−1735 (in Chinese with English abstract). [doi: 10.7544/issn1000-1239.2017.20170178]

[25] Guan PF, Li B, Lv XQ, Zhou JS. Attention enhanced bi-directional LSTM for sentiment analysis. Journal of Chinese Information Processing, 2019,33(2):105−111 (in Chinese with English abstract). http://jcip.cipsc.org.cn/CN/Y2019/V33/I2/105 [doi: CNKI:SUN:MESS.0.2019-02-017]

[26] Zhou X, Wan X, Xiao J. Attention-based LSTM network for cross-lingual sentiment classification. In: Proc. of the 2016 Conf. on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2016. 247−256. [doi: 10.18653/v1/D16-1024]

[27] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser A, Polosukhin I. Attention is all you need. In: Proc. of the Advances in Neural Information Processing Systems. 2017. 5998−6008.

[28] Lin Z, Feng M, Santos CND, Yu M, Xiang B, Zhou B, Bengio Y. A structured self-attentive sentence embedding. arXiv preprint arXiv:1703.03130, 2017.

[29] Wang Y, Sun A, Han J, Liu Y, Zhu X. Sentiment analysis by capsules. In: Proc. of the 2018 Int'l Conf. on World Wide Web. Steering Committee, 2018. 1165−1174. [doi: 10.1145/3178876.3186015]

[30] Schuster M, Paliwal KK. Bidirectional recurrent neural networks. IEEE Trans. on Signal Processing, 1997,45(11):2673−2681.

[31] Ba JL, Kiros JR, Hinton GE. Layer normalization. arXiv preprint arXiv:1607.06450, 2016.

[32] Le Q, Mikolov T. Distributed representations of sentences and documents. In: Proc. of the Int'l Conf. on Machine Learning. JMLR: Workshop&CP, 2014. 1188−1196.

[33] Tang D, Qin B, Liu T. Learning semantic representations of users and products for document level sentiment classification. In: Proc. of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th Int'l Joint Conf. on Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2015. 1014−1023. [doi: 10.3115/v1/P15-1098]

[34] Xu J, Chen D, Qiu X, Huang X. Cached long short-term memory neural networks for document-level sentiment classification. In: Proc. of the 2016 Conf. on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2016. 1660−1669. [doi: 10.18653/v1/D16-1172]

[35] Chen H, Sun M, Tu C, Lin Y, Liu Z. Neural sentiment classification with user and product attention. In: Proc. of the 2016 Conf. on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2016. 1650−1659. [doi: 10.18653/v1/D16-1171]

[36] Pennington J, Socher R, Manning C. Glove: Global vectors for word representation. In: Proc. of the 2014 Conf. on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg: Association for Computational Linguistics, 2014. 1532−1543. [doi: 10.3115/v1/D14-1162]

[37] Liu Y, Bi JW, Fan ZP. A method for multi-class sentiment classification based on an improved one-vs-one (OVO) strategy and the support vector machine (SVM) algorithm. Information Sciences, 2017,394:38−52. [doi: 10.1016/j.ins.2017.02.016]

[38] Zhao W, Ye J, Yang M, Lei Z, Zhang S, Zhao Z. Investigating capsule networks with dynamic routing for text classification. In: Proc. of the 2018 Conf. on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2018. 3110−3119. [doi: 10.18653/v1/D18-1350]

[39] Greff K, Srivastava RK, Koutník J, Steunebrink BR, Schmidhuber J. LSTM: A search space odyssey. IEEE Trans. on Neural Networks and Learning Systems, 2017,28(10):2222−2232. [doi: 10.1109/TNNLS.2016.2582924]

附中文参考文献:

[8] 陈钊,徐睿峰,桂林,陆勤.结合卷积神经网络和词语情感序列特征的中文情感分析.中文信息学报,2015,29(6):172−178. http://jcip.cipsc.org.cn/CN/Y2015/V29/I6/172 [doi: CNKI:SUN:MESS.0.2015-06-024]

[10] 裴颂文,王露露.基于注意力机制的文本情感倾向性研究.计算机工程与科学,2019,41(2):344−353. [doi: CNKI:SUN:JSJK.0.2019-02-023]

[13] 黄发良,冯时,王大玲,于戈.基于多特征融合的微博主题情感挖掘.计算机学报,2017,40(4):872−888. [doi: 10.11897/SP.J.1016.2017.00872]

[14] 黄发良,于戈,张继连,李超雄,元昌安,卢景丽.基于社交关系的微博主题情感挖掘.软件学报,2017,28(3):694−707. http://www.jos.org.cn/1000-9825/5157.htm [doi: 10.13328/j.cnki.jos.005157]

[24] 梁斌,刘全,徐进,周倩,章鹏.基于多注意力卷积神经网络的特定目标情感分析.计算机研究与发展,2017,54(8):1724−1735. [doi: 10.7544/issn1000-1239.2017.20170178]

[25] 关鹏飞,李宝安,吕学强,周建设.注意力增强的双向 LSTM 情感分析.中文信息学报,2019,33(2):105−111. http://jcip.cipsc.org.cn/CN/Y2019/V33/I2/105 [doi: CNKI:SUN:MESS.0.2019-02-017]

李卫疆(1969－),男,博士,教授,主要研究领域为自然语言处理,信息检索.

余正涛(1970－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为自然语言处理,机器翻译,信息检索.

漆芳(1994－),女,硕士,主要研究领域为自然语言处理,情感分析.