结合 BERT 语义融合和关键词特征提取的方面级情感分类研究*

胡耀庭，韩雨桥，石宇航，高　宣，彭玉青
（河北工业大学　人工智能与数据科学学院，天津　300401）

摘　要：方面级情感分类旨在确定句子中给定方面词的情感极性。该任务先前提出的方法无法提取语义信息丰富的上下文初始表示向量，同时也不能精确地捕获局部关键特征的范围。因此，提出了一种结合 BERT 语义融合（BERTSF）和关键词特征提取（KFE）的方面级情感分类模型（KFE-BERTSF）。BERTSF 通过门控融合函数融合 BERT 编码器的高层语义信息，以提取语义信息更加丰富的上下文初始表示向量。KFE 通过动态阈值划分句子的局部上下文和非局部上下文，并利用句法距离掩码（SDMask）和距离感知注意力（ADA）提取两个区域的局部关键特征。基于三个数据集上的实验结果表明，KFE-BERTSF 取得了比基准模型更好的成绩。

关键词：方面级情感分类；BERT 编码器；关键词特征；局部上下文聚焦
中图分类号：TP391　　**文献标识码：**A　　**DOI：**10.19358/j.issn.2097－1788.2024.11.006
引用格式：胡耀庭，韩雨桥，石宇航，等. 结合 BERT 语义融合和关键词特征提取的方面级情感分类研究［J］. 网络安全与数据治理，2024，43(11)：29－36.

Combining BERT semantic fusion and keyword feature extraction for aspect-level sentiment classification

Hu Yaoting，Han Yuqiao，Shi Yuhang，Gao Xuan，Peng Yuqing
（School of Artificial Intelligence，Hebei University of Technology，Tianjin 300401，China）

Abstract：Aspect-level sentiment classification aims to determine the sentiment polarity of a given aspect term in a sentence. Previous methods for this task fail to extract semantically rich initial contextual representation vectors and cannot precisely capture the range of local key features. Therefore，this paper proposes KFE-BERTSF，an aspect-level sentiment classification model that combines BERT semantic fusion（BERTSF）and keyword feature extraction（KFE）. BERTSF integrates high-level semantic information from the BERT encoder using a gating fusion function to extract semantically richer initial contextual representation vectors. KFE divides the sentence into local and non-local contexts using dynamic thresholds，and employs syntax distance mask（SDMask）and distance-aware attention（ADA）to extract local key features from both regions. Experimental results on three datasets show that KFE-BERTSF outperforms benchmark models.

Key words：aspect-level sentiment classification；BERT encoder；keyword feature；local context focus

0　引言

方面级情感分类是情感分析任务的一个分支，旨在确定句子中特定方面词的情感极性。给定句子“Lots of extra space but the keyboard is ridiculously small.”，方面级情感分类（ASC）的任务是对句子中给定的方面词“space”和“keyboard”，应该可以得出对应的情感极性分别为积极和消极。

之前的研究通过将循环神将网络（RNN）与注意力机制[1]进行结合，可以明显提高此类任务的性能。但是注意力机制很容易受到噪声的影响。局部上下文聚焦机制[2]（LCF）发现方面词自身周围的单词对其情感极性的判别更加重要，通过使用 LCF 捕获局部上下文，并使用动态距离掩码（CDM）或动态距离加权（CDW）可以捕获方面词周围的重要信息。同时，随着预训练模型

*基金项目：河北省自然科学基金（F2021202038）

BERT[3]的出现，对 BERT 结构做出针对性的调整也成为了研究方向。尽管上述方法已经取得了显著的成绩，但是仍然存在以下两点问题。（1）未充分挖掘 BERT 语义信息。Ganesh 等人[4]发现 BERT 自身高层的语义信息已经足够丰富，如何更好地利用这些信息仍有待进一步研究。（2）局部关键特征提取不充分。在使用 LCF 的模型中使用固定阈值划分局部上下文，其范围不够精确；同时在两类上下文中使用的特征提取方法未能很好地提取局部关键信息。

为了解决上述问题，本文提出了一个全新的模型，通过结合 BERT 语义融合和关键词特征提取进行方面级情感分类。一是提出 BERT 语义融合模块，利用门控函数将 BERT 编码器不同层的表示向量进行融合。二是提出关键字特征提取模块，根据句子长度动态确定局部上下文和非局部上下文的范围，并通过句法二次掩码和距离感知注意力提取两类上下文的关键特征。三是引入协同注意力模块，将局部特征、全局特征和方面词特征融合，得到融合局部信息和方面词信息的全局特征。

本文的主要贡献总结如下：

（1）将 BERT 编码器的第 9 ~ 12 层通过针对性设计的门控函数进行语义融合，提取到了包含丰富语义的句子初始的上下文表示向量。

（2）提出了关键字特征提取模块。该模块通过句子长度动态确定局部上下文和非局部上下文，并用句法二次掩码和距离感知注意力来分别提取对应区域的局部关键特征。

（3）设计了全新的协同注意力，通过使用注意力机制融合全局特征、局部特征和方面词特征来获得包含丰富语义信息的全新特征。

1　相关工作

近年来，深度学习在各种 NLP 任务中表现出了优异的性能，也被用于 ASC 任务[5]。Tang 等人[6]提出 TD-LSTM，通过利用方面词将上下文分割为两部分分别对两边的上下文进行编码。Huang 等人[7]提出 AOA，采用注意力模块，学习方面和句子的表征，使模型集中在句子的重要部分。同时，预训练模型 BERT 在众多的 NLP 任务上都取得了十分优秀的成绩，包括 ASC 任务。Karimi 等人[8]设计了 BERT-SUM，通过整合 BERT 第 9 ~ 12 层的表示向量进行情感分析。Zhang 等人[9]提出了 DR-BERT，通过动态加权适配器（DRA）在每个步数中动态的理解每句话的关键信息。随着图卷积神经网络（GCN）的兴起，越来越多的学者开始将其运用到 ASC 上。最先将 GCN 运用到 ASC 任务上的是 Zhang[10]等人提出的 AS-CGCN，通过在句子的依赖树上构建 GCN，利用句法信息和单词依赖对上下文进行建模。Li 等人[11]提出 DualGCN，其运用两个 GCN 模块分别捕获上下文的语法和语义信息。

与之不同的是，Zeng 等人[2]提出 LCF-BERT，采用局部上下文聚焦机制获取与方面词周边的局部上下文，对于非局部上下文采用动态距离加权（CDW）或动态距离掩码（CDM）的方式进行特征提取。Xu 等人[12]提出 DLCF-DCA，在此基础上引入动态阈值并结合依赖簇进行情感分类，得到了更好的效果。但是以上方式上仍存在一些问题，预训练模型 BERT 的语义信息也未得到充分的挖掘。为此，本文提出了 KFE-BERTSF，通过融合 BERT 第 9 ~ 12 层的语义信息得到句子初始的表达向量，并提出根据句子长度更精确地划分局部上下文，同时设计了两种新的方法进行局部关键特征的提取。不仅如此，为了融合全局特征、局部特征和方面词特征，本文还提出一个新的协同注意力模块来增强模型的特征交互能力。

2　KFE-BERTSF 模型

本文提出的 KFE-BERTSF 模型架构如图 2 所示。KFE-BERTSF 由嵌入层、BERT 语义融合、关键词特征提取和协同注意力四个部分组成。在嵌入层通过将两类句子序列输入 BERT，获取对应的表示向量。提取到的 BERT 第 9 ~ 12 层的隐藏层向量通过 BERT 语义融合，获得语义信息更加丰富的局部特征初始表示向量。之后通过关键词特征提取获取局部关键特征，得到新的局部特征表示向量。最后，通过协同注意力将全局特征、局部特征和方面词特征进行融合，将全局特征和局部特征进行结合，再使用 softmax 函数输出获得最终的情感分类极性。

2.1　任务定义

给定一个包含 n 个单词的句子 $W^c = \{w_1^c, w_2^c, \cdots, w_n^c\}$ 和包含 m 个单词的方面词 $W^a = \{w_t^a w_{t+1}^a, \cdots, w_{t+m-1}^a\}$。ASC 的任务目标是构建一个情感极性分类器来预测对应方面词的情感极性，情感极性包括积极、消极和中性三种类型。

2.2　嵌入层

为了将单词转换成携带语义信息的表示，对于一个句子 $W^c = \{w_1^c, w_2^c, \cdots, w_n^c\}$，本文使用预训练模型 BERT 将每个单词映射成嵌入向量 $\boldsymbol{e}_t \in \mathbb{R}^{d_h \times 1}$，其中 d_h 是单词向量的隐藏层数。在 KFE-BERTSF 中，输入序列包括两部分，一部分用于语义融合获取初始的局部特征表示向量，另一部分用于获取全局特征表示向量，对应的 token 序列分别为 S_l：“[CLS]” + text sequence + “[SEP]” 和 S_g：“[CLS]” + text sequence + “[SEP]” + aspect + “[SEP]”，其中“[CLS]”为序列的起始标签，“[SEP]”

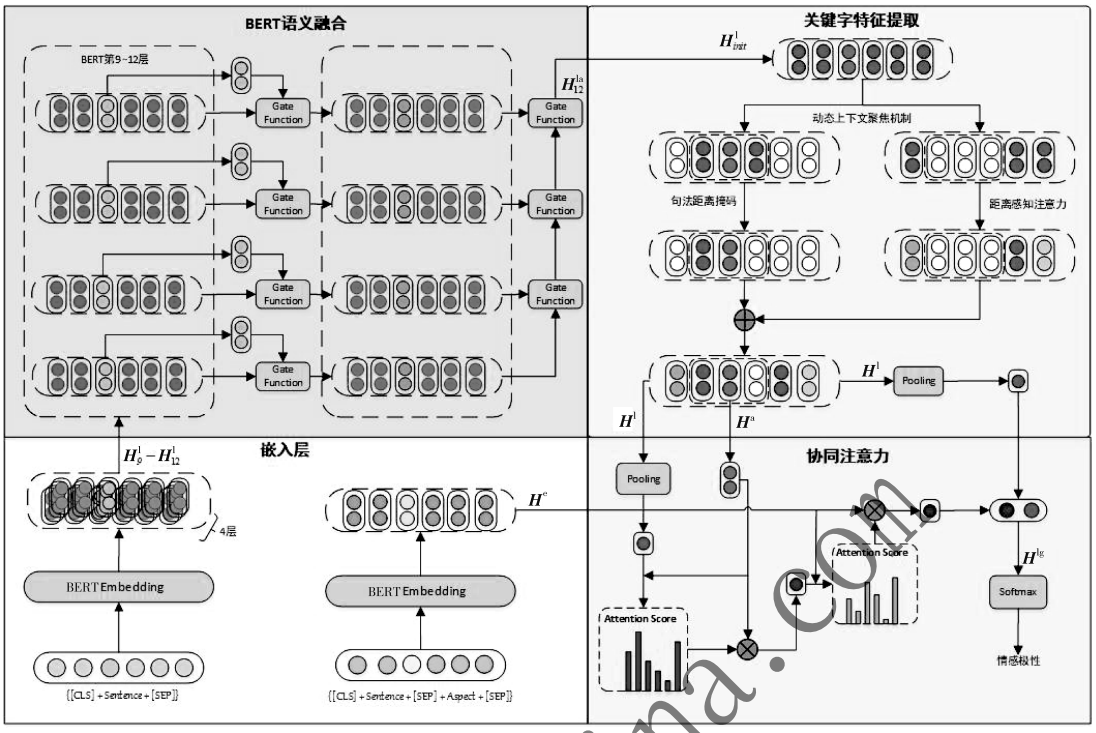

图 1　KFE-BERTSF 模型结构

为序列的结束标签。本文将序列 S_l 和 S_g 分别使用 BERT 编码器进行输入，得到用于提取局部特征和进行语义融合的上下文表示向量 $\boldsymbol{H}_9^l$、$\boldsymbol{H}_{10}^l$、$\boldsymbol{H}_{11}^l$、$\boldsymbol{H}_{12}^l$ 和用于提取全局特征的上下文隐藏向量 $\boldsymbol{H}^g$，如公式（1）（2）所示。

$$\boldsymbol{H}_9^l,\ \boldsymbol{H}_{10}^l,\ \boldsymbol{H}_{11}^l,\ \boldsymbol{H}_{12}^l = \mathrm{BERT}\ (S_l) \tag{1}$$

$$\boldsymbol{H}^g = \mathrm{BERT}\ (S_g) \tag{2}$$

$\boldsymbol{H}_9^l$、$\boldsymbol{H}_{10}^l$、$\boldsymbol{H}_{11}^l$、$\boldsymbol{H}_{12}^l$ 是 token 序列 S_l 经过 BERT 编码器得到的倒数第四层、第三层、第二层和第一层的表示向量。这些向量将会经过门控融合单元形成最终的局部特征初始表示向量 $\boldsymbol{H}_{\mathrm{init}}^l$。

2.3　BERT 语义融合

为了获得包含更多丰富语义信息的局部特征初始表示向量，本文设计了 BERT Semantic Fusion（BERTSF）模块。首先，采用 Peng 等人[13]提出的方法，根据 BERT 编码器得到 BERT 高层的上下文的初始表示向量表示 $\boldsymbol{H}_9^l$、$\boldsymbol{H}_{10}^l$、$\boldsymbol{H}_{11}^l$、$\boldsymbol{H}_{12}^l$，提取对应的方面词表示向量 $\boldsymbol{H}_9^a$、$\boldsymbol{H}_{10}^a$、$\boldsymbol{H}_{11}^a$、$\boldsymbol{H}_{12}^a$。为了增强每层的表示向量与方面词的关联程度，对于第 n 层的上下文表示向量 $\boldsymbol{H}_n^l$，使用门控融合函数 G 将对应层的方面词表示向量 $\boldsymbol{H}_n^a$ 融合进每层的上下文表示向量中，得到隐藏向量 $\boldsymbol{H}_n^{la}$。接着再次使用门控融合方法 G 来融合来自上一层经过语义融合的表示向量 $\boldsymbol{H}_n^{la'}$。最终得到新的局部特征初始表示向量 $\boldsymbol{H}_{\mathrm{init}}^l$。具体的过程如公式（3）~（5）所示。

$$\boldsymbol{H}_{\mathrm{init}}^l = \boldsymbol{H}_{12}^{la'} \tag{3}$$

$$\boldsymbol{H}_n^{la'} = \begin{cases} G\ (\boldsymbol{H}_n^{la},\ \boldsymbol{H}_{n-1}^{la'}), & n = 10,\ 11,\ 12 \\ \boldsymbol{H}_n^{la}, & n = 9 \end{cases} \tag{4}$$

$$\boldsymbol{H}_n^{la} = G\ (\boldsymbol{H}_n^l,\ \boldsymbol{H}_n^a) \tag{5}$$

其中 n 是隐藏向量的层数，$\boldsymbol{H}_n^{la}$ 是经过 G 将第 n 层方面词表示向量与对应层数的上下文表示向量融合后产生的新的上下文表示向量。G 是门控融合函数，利用 Sigmoid 函数计算两向量之间的关联程度 $\boldsymbol{r}_i$，并与需要进行特征融合的向量相乘进行特征融合。G 的计算过程如式（6）、式（7）所示。

$$G\ (\boldsymbol{H}_i,\ \boldsymbol{H}_j) = \boldsymbol{r}_i \odot \boldsymbol{H}_i \tag{6}$$

$$\boldsymbol{r}_i = \mathrm{Sigmoid}\ (\boldsymbol{W}_1 \boldsymbol{H}_i + \boldsymbol{W}_2 \boldsymbol{H}_j) \tag{7}$$

其中 $\boldsymbol{r}_i$ 是向量 $\boldsymbol{H}_i$ 和 $\boldsymbol{H}_j$ 的语义相关性权重，$\boldsymbol{W}_1 \in \mathbb{R}^{d_w \times d_w}$，$\boldsymbol{W}_2 \in \mathbb{R}^{d_w \times d_w}$ 属于可学习的参数，$\boldsymbol{H}_i$ 和 $\boldsymbol{H}_j$ 代表需要进行关联程度计算的两个向量，⊙代表逐元素相乘操作。

2.4　关键词特征提取

本文设计了关键字特征提取（KFE）来充分地提取局部关键特征。通过使用语义相关距离（SemRD）来计算方面词与上下文之间的语义关联程度，并根据句子长度动态捕获句子的局部上下文和非局部上下文的范围。对于两类上下文，分别使用句法距离掩码（SDMask）和距离感知注意力（ADA）提取对应区域的特征。

2.4.1 动态上下文聚焦机制

为了确定局部上下文范围，首先要计算单词的语义相关距离（SemRD）。同时，本文在后续特征提取中也使用了句法距离（SynRD）。当方面词有多个单词组成时，计算与方面词的平均距离得到 SemRD 和 SynRD。LCF 通过式（8）、式（9）计算 SemRD 和 SynRD。

$$\mathrm{SemRD}_i = \left| i - P^{\mathrm{a}}_{\mathrm{sem}} \right| - \left\lfloor \frac{m}{2} \right\rfloor \tag{8}$$

$$\mathrm{SynRD}_i = \left| i - P^{\mathrm{a}}_{\mathrm{syn}} \right| - \left\lfloor \frac{m}{2} \right\rfloor \tag{9}$$

其中 i 是上下文单词的位置，$P^{\mathrm{a}}_{\mathrm{sem}}$ 和 $P^{\mathrm{a}}_{\mathrm{syn}}$ 分别代表语义距离和语法距离中方面词中心词的位置。m 是方面词的序列长度。

本文设计了动态阈值 α_{d} 来计算不同句子的局部上下文范围。同时在后续语法距离掩码中的阈值 α_{y} 参考文献[12]中阈值的设定。α_{d} 和 α_{y} 的计算过程如式（10）、式（11）所示。

$$\alpha_{\mathrm{d}} = n \times \beta + k \tag{10}$$

$$\alpha_{\mathrm{y}} = \log_a(M_{\mathrm{d}}) + a - 1 \tag{11}$$

其中 n 代表句子长度，β、k 和 a 属于自定义参数。M_{d} 代表当前句子的最大句法距离。

2.4.2 句法距离掩码

为了减少局部上下文中噪声的干扰，本文设计了句法距离掩码（SDMask）。SDMask 会再次计算局部上下文中每个单词的 SynRD，处于句法掩码阈值内的，会将其对应的权重向量设置为单位向量，否则设置为零向量，进而计算出对应的权重向量 $\boldsymbol{s}_i$。具体的计算过程如公式（12）所示。

$$\boldsymbol{s}_i = \begin{cases} \boldsymbol{O}, & \mathrm{SynRD}_i > \alpha_{\mathrm{y}} \\ \boldsymbol{E}, & \mathrm{SynRD}_i \leqslant \alpha_{\mathrm{y}} \end{cases} \tag{12}$$

其中 i 代表单词的位置，SynRD_i 为第 i 个单词的的语法相关距离，$\boldsymbol{E} \in \mathbb{R}^{d_w}$ 为单位向量，$\boldsymbol{O} \in \mathbb{R}^{d_w}$ 为零向量，$\boldsymbol{s}_i$ 为该单词的权重向量。

2.4.3 距离感知注意力

为防止模型忽略处于非局部上下文中的关键单词特征，本文提出了距离感知注意力（ADA），ADA 会将得分函数 f 计算出的相关程度权重向量和距离权重向量 $\boldsymbol{d}_i$ 相加，并利用 softmax 函数计算得到该单词的最终的权重向量 $\boldsymbol{a}_i$，$\boldsymbol{a}_i \in (0, 1)$。式（13）~式（16）展示了具体的计算过程。

$$\boldsymbol{h}^{\mathrm{a}}_{\mathrm{pool}} = \mathrm{Pool}(\boldsymbol{H}^{\mathrm{a}}) \tag{13}$$

$$\boldsymbol{a}_i = \mathrm{softmax}(f(\boldsymbol{h}^{\mathrm{a}}_{\mathrm{pool}}, \boldsymbol{h}^{\mathrm{l}}_i) + \boldsymbol{d}_i) \tag{14}$$

$$f(\boldsymbol{h}^{\mathrm{a}}_{\mathrm{pool}}, \boldsymbol{h}^{\mathrm{l}}_i) = \tanh(\boldsymbol{h}^{\mathrm{a}}_{\mathrm{pool}} \boldsymbol{W}_3 \boldsymbol{h}^{\mathrm{l}}_i) \tag{15}$$

$$\boldsymbol{d}_i = \left(1 - \frac{\mathrm{SemRD}_i - \alpha_{\mathrm{d}}}{n}\right) \cdot \boldsymbol{E} \tag{16}$$

其中，$\boldsymbol{H}^{a}$ 为从 $\boldsymbol{H}^{l}_{\mathrm{init}}$ 中提取出的方面词表示向量，$\boldsymbol{h}^{\mathrm{a}}_{\mathrm{pool}}$ 是 $\boldsymbol{H}^{\mathrm{a}}$ 经过平均池化后得到的方面词隐藏向量，i 代表单词位置，$\boldsymbol{h}^{\mathrm{l}}_i$ 是 $\boldsymbol{H}^{\mathrm{l}}_{\mathrm{init}}$ 第 i 个单词的隐藏向量表示，$\boldsymbol{W}_3 \in \mathbb{R}^{1 \times d_w}$ 为可学习参数，$\boldsymbol{d}_i$ 是 $\boldsymbol{H}^{\mathrm{l}}_{\mathrm{init}}$ 第 i 个单词的距离权重向量，n 为句子长度，$\boldsymbol{E} \in \mathbb{R}^{d_w}$ 为单位向量。

2.4.4 局部关键特征提取

对于使用动态阈值划分出的局部上下文和非局部上下文，分别使用 SDMask 和 ADA 计算出对应单词的权重向量 $\boldsymbol{s}_i$ 和 $\boldsymbol{a}_i$。最终将所有单词的权重向量拼接成权重矩阵 $\boldsymbol{M}$，与局部特征初始表示向量 $\boldsymbol{H}^{\mathrm{l}}_{\mathrm{init}}$ 相乘得到局部关键特征表示向量 $\boldsymbol{H}^{\mathrm{l}}$。具体的计算过程如式（17）~式（19）所示。

$$\boldsymbol{v}_i = \begin{cases} \boldsymbol{s}_i, & \mathrm{SemRD}_i > \alpha_{\mathrm{d}} \\ \boldsymbol{a}_i, & \mathrm{SemRD}_i \leqslant \alpha_{\mathrm{d}} \end{cases} \tag{17}$$

$$\boldsymbol{M} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_n] \tag{18}$$

$$\boldsymbol{H}^{\mathrm{l}} = \boldsymbol{H}^{\mathrm{l}}_{\mathrm{init}} \odot \boldsymbol{M} \tag{19}$$

其中 α_{d} 是计算局部上下文的阈值，i 是单词的位置，$\boldsymbol{v}_i$ 是每个单词对应的权重向量，$\boldsymbol{M}$ 是由 $\boldsymbol{v}_i$ 组合成的局部特征初始表示向量 $\boldsymbol{H}^{\mathrm{l}}_{\mathrm{init}}$ 的权重矩阵，n 为输入序列的长度，$\odot$ 代表按位相乘操作。

2.5 协同注意力

为了尽可能减少全局特征表示向量 $\boldsymbol{H}^{\mathrm{g}}$ 中的噪声，并保留全局特征中比较重要的单词特征，本文提出了协同注意力（Co-Attention）。首先，本文使用注意力机制对 $\boldsymbol{H}^{\mathrm{l}}$ 的平均池化向量 $\boldsymbol{H}^{\mathrm{l}}_{\mathrm{pool}}$ 与方面词特征向量进行计算，如式（20）~式（23）所示。

$$\boldsymbol{H}^{\mathrm{l}}_{\mathrm{pool}} = \mathrm{Pool}(\boldsymbol{H}^{\mathrm{l}}) \tag{20}$$

$$f(\boldsymbol{H}^{\mathrm{l}}_{\mathrm{pool}}, \boldsymbol{h}^{\mathrm{a}}_i) = \tanh(\boldsymbol{H}^{\mathrm{l}}_{\mathrm{pool}} \boldsymbol{W}_5 \boldsymbol{h}^{\mathrm{a}}_i) \tag{21}$$

$$\boldsymbol{\theta}_i = \frac{\exp(f(\boldsymbol{H}^{\mathrm{l}}_{\mathrm{pool}}, \boldsymbol{h}^{\mathrm{a}}_i))}{\sum_{i=1}^{m} \exp(f(\boldsymbol{H}^{\mathrm{l}}_{\mathrm{pool}}, \boldsymbol{h}^{\mathrm{a}}_i))} \tag{22}$$

$$\boldsymbol{H}^{\mathrm{a}'} = \sum_{i=1}^{m} \boldsymbol{h}^{\mathrm{a}}_i \boldsymbol{\theta}_i \tag{23}$$

其中 f 为得分函数，$\boldsymbol{h}^{\mathrm{a}}_i$ 是 $\boldsymbol{H}^{\mathrm{a}}$ 中第 i 个单词的方面词表示向量，$\boldsymbol{W}_5 \in \mathbb{R}^{1 \times d_w}$ 为可学习参数，m 为方面词长度，$\boldsymbol{\theta}_i$ 为注意力权重向量，最后得到方面词特征的表示向量 $\boldsymbol{H}^{a'}$。类似地，本文将得到的方面词特征表示向量融合进全局特征表示向量中，得到全局特征表示向量 $\boldsymbol{H}^{g'}$。具体的计算过程如式（24）~式（26）所示。

$$f(\boldsymbol{H}^{\mathrm{a}'}, \boldsymbol{h}^{\mathrm{g}}_j) = \tanh(\boldsymbol{H}^{\mathrm{a}'} \cdot \boldsymbol{W}_6 \cdot \boldsymbol{h}^{\mathrm{g}}_j) \tag{24}$$

$$\boldsymbol{\lambda}_j = \frac{\exp(f(\boldsymbol{H}^{\mathrm{a}'}, \boldsymbol{h}^{\mathrm{g}}_j))}{\sum_{j=1}^{n} \exp(f(\boldsymbol{H}^{\mathrm{a}'}, \boldsymbol{h}^{\mathrm{g}}_j))} \tag{25}$$

$$\boldsymbol{H}^{g'} = \sum_{j=1}^{n} \boldsymbol{h}_j^g \boldsymbol{\lambda}_j \tag{26}$$

$\boldsymbol{h}_j^g$ 是 $\boldsymbol{H}^g$ 中第 j 个单词的全局特征表示向量，$\boldsymbol{W}_6 \in \mathbb{R}^{1\times d_u}$ 为可学习参数，n 为输入序列长度，$\boldsymbol{\lambda}_j$ 是注意力权重向量，$\boldsymbol{H}^{g'}$ 为表示全局特征的注意力表示向量。

2.6 分类层

本文将池化后的局部关键特征表示向量 $\boldsymbol{H}_{\text{pool}}^l$ 和经过融合后的全局特征表示向量 $\boldsymbol{H}^{g'}$ 进行拼接，得到最终的表示向量 $\boldsymbol{H}^{lg}$。最后通过 softmax 函数生成情感极性概率分布 p。计算的过程如式（27）、式（28）所示。其中 $\boldsymbol{W}_7 \in \mathbb{R}^{d_k\times d_k}$，$\boldsymbol{b} \in \mathbb{R}^{d_k}$ 是可学习的权重和偏置。

$$\boldsymbol{H}^{lg} = [\boldsymbol{H}_{\text{pool}}^l ; \boldsymbol{H}^{g'}] \tag{27}$$

$$p = \text{softmax}\ (\boldsymbol{W}_7^{\mathrm{T}} \boldsymbol{H}^{lg} + \boldsymbol{b}) \tag{28}$$

3 实验

3.1 数据集及参数设置

本文在 Restaurant、Laptop 和 Twitter 数据集上对模型进行了评估。Restaurant 和 Laptop 数据集来自 2014 年 SemEval－2014[14] 上公开的用于 ABSA 任务的数据集。Twitter 数据集来自 tweets 评论[15]。所有的数据集都包含三种情感极性：积极、消极和中性。三种数据集的数据统计信息如表 1 所示。

本文对实验的超参数进行了如下的设置：batch_size 为 16，epoch 设为 10，BERT 隐藏向量的初始维度设为 768。采用 Adam 优化器，模型的权重和偏差使用 Xavier 均匀分布进行初始化。对于三个数据集，学习率设置为 2×10^{-5}，L_2 正则化系数设为 1×10^{-5}，dopout 设置为 0.1。用于调整阈值的超参数 β 设置为 0.05，k 设置为 1.5，a 设置为 2。

表 1 数据集统计信息

数据集	积极		中性		消极	
	训练数据	测试数据	训练数据	测试数据	训练数据	测试数据
Restaurant	2 164	728	637	196	807	196
Laptop	994	341	464	169	870	128
Twitter	1 561	173	3127	346	1 560	173

3.2 对比实验

本文使用分类准确率（Accuracy）和 Macro-$F1$ metrics 作为评价指标。为了全面评估模型的性能，本文将 KFE-BERTSF 与许多基准模型进行了比较，表 2 展示了比较结果。用于对比的模型介绍如下：

IAN[16] 设计了交互式的注意力机制来分别生成方面词和上下文的表示向量。

RAM[17] 使用多个注意力结构和记忆网络学习句子的表示向量。

AEN-BERT[18] 使用 BERT 编码器获取上下文和方面词的编码表示向量，并利用多头自注意力机制建模获取上下文和方面词之间的关系。

BERT[3] 通过将句子对和方面词输入到 BERT 模型并使用［CLS］标签获得句子的最终特征向量。

BERT-SUM[8] 将 BERT 提取的第 9 到第 12 层表示向量进行加和平均来获得句子的最终表示向量。

DR-BERT[9] 设计了动态权重适配器（DRA）对句子的理解进行动态变更，获取与方面词相关的最关键的信息。

LCF-BERT[2] 设计了局部上下文动态机制，通过计算上下文单词与方面词的语义距离判断其重要程度，并使用上下文动态加权机制（CDW）对不在局部上下文的单词进行加权。

LCFS-BERT[19] 在 LCF-BERT 的基础上对上下文单词与方面词的距离计算做了更改，通过使用句法距离计算两者之间的重要程度。

DLCF-DCA[12] 对 LCF 设计了动态阈值，并结合依赖簇构建上下文的表示向量。

表 2 展示了本文的模型与基准模型的性能比较结果。相比于基准模型，KFE-BERTSF 在 Laptop、Restaurant 和 Twitter 数据集上准确率分别提升了 0.94%、0.92%、0.41%，$F1$ 值提升了 0.54%、0.66%、0.31%，比基准模型表现更加优秀，其原因可能包括以下几点：一是融合了 BERT 编码器第 9～12 层的语义信息获得更加丰富的上下文初始表示向量，解决了上述 BERT Models 挖掘 BERT 语义信息不足的问题；二是捕获了范围更加精确的局部上下文并使用 SDMask 和 ADA 分别提取对应区域的特征，解决了 LCF Models 局部上下文范围不精确和特征提取不充分的问题，也解决了 Attention Models 容易引入噪声的问题；三是使用了全新的协同注意力机制，可以更好地融合局部特征，全局特征和方面词特征，解决了特征契合度不充分的问题。

3.3 消融实验

为了分析 KFE-BERTSF 模型每个模块的重要性，本文设计了消融实验，实验变量在下面展示，实验结果如表 3 所示。其中：w/odlcf 表示不使用 KFE 中的动态阈值算法，采用固定阈值确定局部上下文范围；w/osdmask&ada 表示不使用 KFE 中的句法距离掩码和距离感知注意力进行特征提取，直接使用 CDW 进行局部特征提取。

通过表 3 可以看出在准确率和 Macro-*F*1 方面，各个消融模型的性能均无法与 KFE-BERTSF 模型相比。与 KFE-BERTSF 模型相比，不使用 BERTSF 实验的准确度和 *F*1 值在三个数据集上都有降低，尤其在 Twitter 数据集上降低的更为明显，这说明该模块可以更加深入地提取类似 Twitter 数据集里抽象的语义信息，增加了模型的语义理解能力。同理，去掉 KFE 模块和 Co-Attention 模块模型性能下降明显，说明 KFE 模块可以增加模型提取关键特征的能力，Co-Attention 模块可以更好地融合各个部分的特征，可以提高模型的性能。同时，去掉 Co-Attention 模块后的模型性能下降最严重，证明融合方法对模型性能的影响很大。本文在 KFE 模块中单独进行了消融实验，‘w/o dlcf’指不使用动态阈值划分局部上下文，‘w/o sdmask&ada’指不使用句法距离掩码和距离感知注意力进行特征提取，这些实验结果相比于原模型性能均有不同程度的下降，证明了使用动态阈值捕获局部上下文范围，并使用句法距离掩码和距离感知注意力可以更好地提取局部关键特征。

表 2　KFE-BERTSF 与基准模型的比较结果（%）

Models		Laptop		Restaurant		Twitter	
		Accuracy	*F*1	Accuracy	*F*1	Accuracy	*F*1
Attention Models	IAN	72. 10	—	78. 60	—	—	—
	RAM	74. 49	71. 35	80. 23	70. 80	69. 36	67. 30
	AEN-BERT	79. 93	76. 31	83. 12	73. 76	74. 71	73. 13
BERT Models	BERT *	79. 15	74. 70	85. 00	77. 85	74. 57	73. 57
	BERT-SUM	79. 55	76. 81	86. 30	79. 68	—	—
	DR-BERT *	81. 36	77. 59	87. 03	81. 42	76. 32	74. 83
LCF Models	LCF-BERT *	80. 41	76. 72	86. 34	80. 39	72. 83	70. 50
	LCFS-BERT	80. 52	77. 13	86. 71	80. 31	—	—
	DLCF-DCA *	80. 82	77. 67	85. 68	79. 60	—	—
Ours	KFE-BERTSF	82. 13	78. 57	87. 95	82. 08	76. 73	75. 19

注：本文根据论文中发表的源代码复现的模型结果用“*”表示。所有的超参数和实验环境都严格遵循源文件。“-”代表未报道的实验结果。实验结果为多组实验结果的平均值。

表 3　KFE-BERTSF 的消融实验（%）

BERTSF	KFE	Co-Attention	Laptop		Restaurant		Twitter	
			Accuracy	*F*1	Accuracy	*F*1	Accuracy	*F*1
√	√	√	82. 13	78. 57	87. 95	82. 08	76. 73	75. 19
×	√	√	81. 7	78. 21	87. 68	81. 41	75. 14	73. 65
√	×	√	80. 41	76. 45	86. 79	79. 4	75	73. 65
√	w/o dlcf	√	80. 72	77. 75	86. 96	81. 14	76. 68	75. 03
√	w/o sdmask&ada	√	81. 03	76. 59	87. 14	80. 33	75. 14	73. 76
√	√	×	80. 88	77. 5	86. 7	79. 11	74. 86	73. 19

注：“BERTSF”指 BERT 语义融合，“KFE”指关键字特征提取，“Co-Attention”指协同注意力。“√”和“×”表示是否使用了该模块，“w/o”表示未使用该模块中的指定方法。

3. 4　BERT 层数的影响

为了探究 BERTSF 模块中融合的 BERT 编码器表示向量的层数对模型性能的影响，本文在 Laptop、Restaurant 和 Twitter 三个数据集上进行了实验，结果如图 2 所示。可以看到当融合 BERT 第 9 ~ 12 层时模型的实验结果达到最佳。同时可知，当融合的层数过多时，会引入更多的噪声，当融合的层数过少时，融合的语义特征会变少，进而影响模型性能。只有当融合的层数适当时，融合的语义信息充分且噪声量适中。

3. 5　局部上下文阈值的影响

KFE-BERTSF 中局部上下文的范围与确定动态阈值的参数 β 和 k 有关，为了探究参数变化对模型性能的影响，

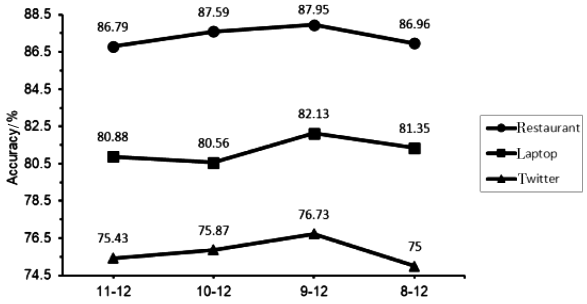

图 2　融合 BERT 不同的层数对模型 Accuracy 的影响

本文设定了不同的 β 和 k。实验结果如图 3、图 4 所示。本文在 Laptop、Restaurant 和 Twitter 三个数据集上进行了实验，对于探究 β 影响的实验，本文将 k 设置为 0；同样，对于探究 k 影响的实验，本文将 β 设置为 1.5。通过实验结果可知，当 β 或 k 超出或低于最佳参数值后，模型的性能会下降，证明了精确捕获局部上下文范围对于模型的性能有着重要的影响。

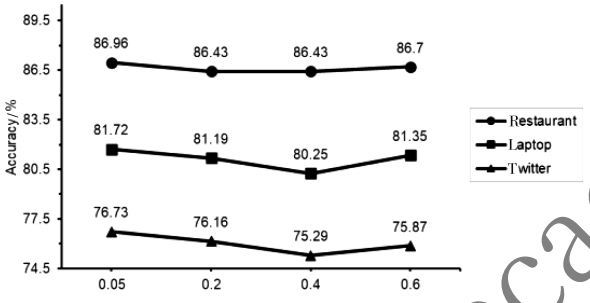

图 3　参数 β 对模型 Accuracy 的影响

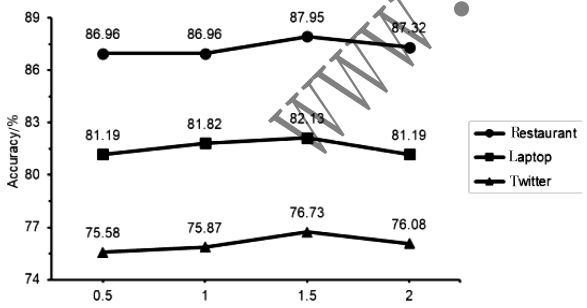

图 4　参数 k 对模型 Accuracy 的影响

3.6　案例分析

为了更好地了解 KFE-BERTSF 中主要模块的作用，本文从 Restaurant 数据集中选取了一个样本进行案例分析，其为“The staff should be a bit more friendly.”。本文对样本的注意力得分进行了可视化，如图 5 所示，展现了模型提取局部关键特征、全局特征和综合特征的效果。图（a）展示了样本经过 KFE 后各个单词的权重，通过热力图可以看到，经过此模块后局部上下文的大部分单词

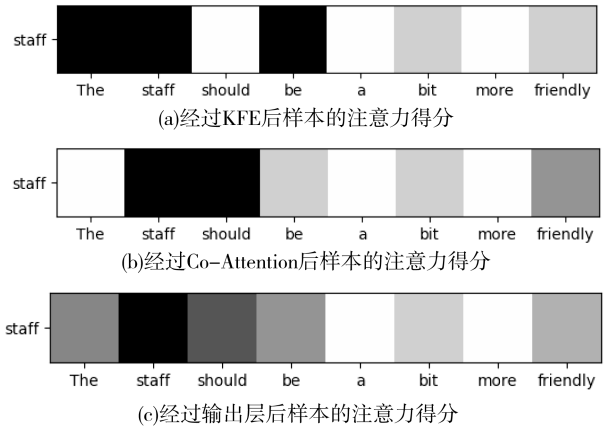

图 5　样本的注意力得分

和距离方面词较远的重要观点词被保留。图（b）展示了样本经过 Co-Attention 后，对于 KFE 中缺少的特征，例如单词“should”，经过 Co-Attention 模块的处理，“should”的权重明显增大。图（c）展示了局部特征和全局特征结合后的权重热力图，结合后的综合特征不仅保留了来自各自特征的优势，同时也弥补了各自的不足之处。通过分析证明，KFE-BERTSF 可以很好地捕获到重要的观点词，提高了模型的性能。

4　结论

本文提出了融合了局部关键特征和全局特征并用于方面级情感分析的模型 KFE-BERTSF。为了提取包含更多语义信息的上下文初始表示向量，本文提出了 BERT 层语义融合方法，通过门控融合函数对 BERT 第 9～12 层的表示向量进行融合。同时，提出关键字特征提取，改进局部上下文聚焦机制，并且提出不同的特征提取方法获取局部关键特征。利用 Co-Attention 更好地融合了局部特征、全局特征和方面词特征。实验表明，KFE-BERTSF 在多个数据集的表现均优于其他基准模型。

参考文献

[1] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [J]. arXiv. 1706. 03762, 2017.

[2] ZENG B, YANG H, XU R, et al. LCF: a local context focus mechanism for aspect-based sentiment classification [J]. Applied Sciences, 2019, 9 (16): 3389.

[3] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding [J]. arXiv: 1810. 04805, 2018.

[4] JAWAHAR G, SAGOT B, DJAMÉ SEDDAH. What does BERT learn about the structure of language? [C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019.

[5] PARK H J, SONG M, SHIN K S. Deep learning models and datasets for aspect term sentiment classification: implementing holistic recurrent attention on target-dependent memories [J]. Knowledge-Based Systems, 2020, 187 (Jan.): 104825.

[6] TANG D, QIN B, FENG X, et al. Effective LSTMs for target-dependent sentiment classification [J]. arXiv: 1512.01100, 2015.

[7] HUANG B, OU Y, CARLEY K M. Aspect level sentiment classification with attention-over-attention neural networks [J]. arXiv: 1804.06536, 2018.

[8] KARIMI A, ROSSI L, PRATI A. Improving BERT performance for aspect-based sentiment analysis [J]. arXiv: 2010.11731, 2020.

[9] ZHANG K, ZHANG K, ZHANG M, et al. Incorporating dynamic semantics into pre-trained language model for aspect-based sentiment analysis [J]. arXiv: 2203.16369, 2022.

[10] ZHANG C, LI Q, SONG D. Aspect-based sentiment classification with aspect-specific graph convolutional networks [J]. arXiv: 1909.03477, 2019.

[11] LI R, CHEN H, FENG F, et al. Dual graph convolutional networks for aspect-based sentiment analysis [C]//Proceedings of 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2021.

[12] Studies fromsouth China normal university provide new data on neural computation (combining dynamic local context focus and dependency cluster attention for aspect-level sentiment classification) [J]. Robotics & Machine Learning Daily News, 2022 (6): 12 - 13.

[13] YUQING P, TENGFEI X, HONGTAO Y. Cooperative gating network based on a single BERT encoder for aspect term sentiment analysis [J]. Applied Intelligence: The International Journal of Artificial Intelligence, Neural Networks, and Complex Problem-Solving Technologies, 2022 (5): 52.

[14] PONTIKI M, GALANIS D, PAVLOPOULOS J, et al. SemEval-2014 task 4: aspect based sentiment analysis [C]//Proceedings of International Workshop on Semantic Evaluation, 2014.

[15] DONG L, WEI F, TAN C, et al. Adaptive recursive neural network for target. dependent Twitter sentiment classification [C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 2014, 2: 49 - 54.

[16] MA D, LI S, ZHANG X, et al. Interactive attention networks for aspect-level sentiment classification [J]. arXiv: 1709.00893, 2017.

[17] CHEN P, SUN Z, BING L, et al. Recurrent attention network on memory for aspect sentiment analysis [C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017: 452 - 461.

[18] SONG Y, WANG J, JIANG T, et al. Attentional encoder network for targeted sentiment classification [J]. arXiv: 1902.09314, 2019.

[19] PHAN M H, OGUNBONA P O. Modelling context and syntactical features for aspect-based sentiment analysis [C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020: 3211 - 3220.

（收稿日期：2024 - 08 - 12）

作者简介：

胡耀庭（1999 - ），男，硕士研究生，主要研究方向：情感分析、自然语言处理。

彭玉青（1969 - ），通信作者，女，硕士，教授，主要研究方向：情感分析、自然语言处理、计算机视觉。E-mail: pengyuqing@ hebut. edu. cn。

版权声明

凡《网络安全与数据治理》录用的文章，如作者没有关于汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权等版权的特殊声明，即视作该文章署名作者同意将该文章的汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权授予本刊，本刊有权授权本刊合作数据库、合作媒体等合作伙伴使用。同时，本刊支付的稿酬已包含上述使用的费用，特此声明。

《网络安全与数据治理》编辑部