基于 BiLSTM-CNN 的文本情感分类

学号：19124418　姓名：纪斌斌

1 研究问题

1.1 研究背景

近年来，互联网用户普及度高，人们越来越多在社交媒体发表自己的观点和评论。因此，网络上涌现出大量的文本数据。文本情感分类是情感分析中的一个核心问题，通过对文本进行情感倾向性的分类可以帮助消费者、商家或者政府做出选择或者决策，具有重要的商业价值和社会意义[1]。

1.2 研究问题特点及难点

文本数据多呈现非结构化、长短不一的特点，并且人类的自然语言情感表达复杂，常伴有双重含义、反讽、玩笑等情绪[2]。因此，文本情感分类的难点主要集中于文本特征提取和情感语义的机器理解。具体表现为文本数据规范而有效的向量化表达，结合上下文和关键词的语义理解以及局部特征和全局特征的提取。

1.3 研究内容

本研究关注于多类别的文本情感分类，对比分析了 TextCNN 模型、融合注意力机制和文本变长处理的 BiLSTM 模型、融合注意力机制的 CNN-BiLSTM 模型和 BiLSTM-CNN 四种模型的分类能力，结果显示 BiLSTM-CNN 模型效果最佳。该模型首先使用 BiLSTM 结合上下文为文本生成新的特征向量，并在 BiLSTM 后接 CNN 进行局部特征提取和语义理解。最后，使用查准率、召回率、F1-score 三个指标对模型分类结果进行评价[3]。

2 研究数据

2.1 数据来源

数据选自 Kaggle “Emotions dataset for NLP classification tasks” 数据集。

2.2 数据基本信息

数据中包含训练、验证、测试三个数据集，每条数据由文本及其对应情感分类组成。数据中共包含六类情感：sadness，anger，love，surprise，fear，joy。表 1 统计了三个数据集的基本信息。

表 1 数据基本信息

数据集	样本数	标签类别
训练集	16000	6
验证集	2000	6
测试集	2000	6

2.3 数据预处理

首先将文本和标签分离。其次对文本进行大小写统一、缩写更正、去停词、符号删除以及分词操作。最后对标签进行数字编码。数据预处理前后对比见表 2。

表 2 数据预处理前后对比

数据预处理前后对比	
处理前的文本	im grabbing a minute to post, i feel greedy wrong.
处理后的文本	[grabbing, minute, post, feel, greedy, wrong]
处理前的标签	anger
处理后的标签	0

3 解决方案介绍

3.1 BiLSTM-CNN 基本原理

CNN 的本质在于构建多个能够提取数据特征的滤波器，通过对输入数据进行逐层卷积和池化操作来提取数据之间隐藏的拓扑结构特征。随着层数的不断增加，提取的特征越来越抽象，最后将这些抽象特征通过全连接层汇合，并通过 softmax 或 sigmoid 激活函数解决分类问题和回归问题。CNN 的特点之一在于可提取输入数据的局部特征，并逐层组合抽象生成高层特征，有效实现故障诊断与识别。

BiLSTM 能够有效地处理较长的文本数据，提取词语与上下文之间的关系，对于词语语序更加敏感。每一个 LSTM 的神经单元是由细胞状态，输入门，遗忘门，输出门三个门所组成，模型通过遗忘门决定从细胞状态中丢弃什么信息，通过记忆门选择要记忆的信息，通过输出门控制当前状态的输出信息。BiLSTM 由一个正向的 LSTM 和一个逆向的 LTSM 所组成，能够完成长距离的上下文理解。

下式为 LSTM 各单元控制逻辑公式。设输入时间序列为x，t为当前时刻，$g(t)$表示状态输入单元；$h(t)$ 表示状态输出单元；$M(t)$ 表示记忆单元; $i(t)$、$o(t)$、$f(t)$分别表示输入门限、输出门限以及遗忘门限。其中, σ 表示 sigmoid 激活函数[4]。

$$g(t) = \tanh\big(W_{xg}g(t-1)\big) + W_{hg}h(t-1) + b_g \tag{1}$$

$$i(t) = \sigma(W_{xi}i(t-1) + W_{hi}h(t-1) + b_i \tag{2}$$

$$f(t) = \sigma(W_{xf}i(t-1) + W_{hf}h(t-1) + b_f) \tag{3}$$

$$o(t) = \sigma(W_{xo}i(t-1) + W_{ho}h(t-1) + b_o) \tag{4}$$

$$M(t) = f(t)M(t-1) + i(t)g(t) \tag{5}$$

$$h(t) = o(t)\tanh\,(M(t)) \tag{6}$$

3.2 BiLSTM-CNN 模型构建

CNN 网络在提取局部特征方面具有优势，BiLSTM 对全局文本结合上下文的特征提取具有优势。故考虑将两种模型进行结合，该模型首先利用 BiLSTM 层接收句子中每一个词语的 word embedding，并输出储存前后文信息的特征，然后将 BiLSTM 层的输出紧接着输入 CNN 提取局部特征。最后利用全连接完成文本情感多分类问题。图 1 为该模型的结构。

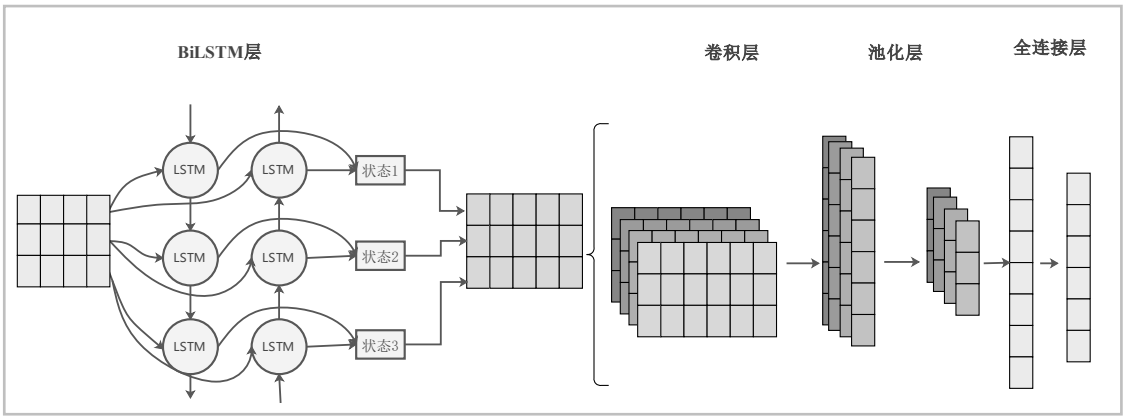

图 1 BiLSTM-CNN 模型结构

4 实验设计

4.1 实验对比模型

4.1.1 TextCNN

TextCNN 使用多个不同尺寸的卷积核对输入数据进行一维卷积提取句子中的局部特征，并对卷积结果进行最大池化实现关键特征提取。相较于 BiLSTM-CNN，TextCNN 网络结构简单且训练速度快，但也出现长距离特征提取受限以及对语序不敏感的问题。如图 2 所示，本文构建的 TextCNN 由嵌入层、卷积层、池化层以及全连接层组成。

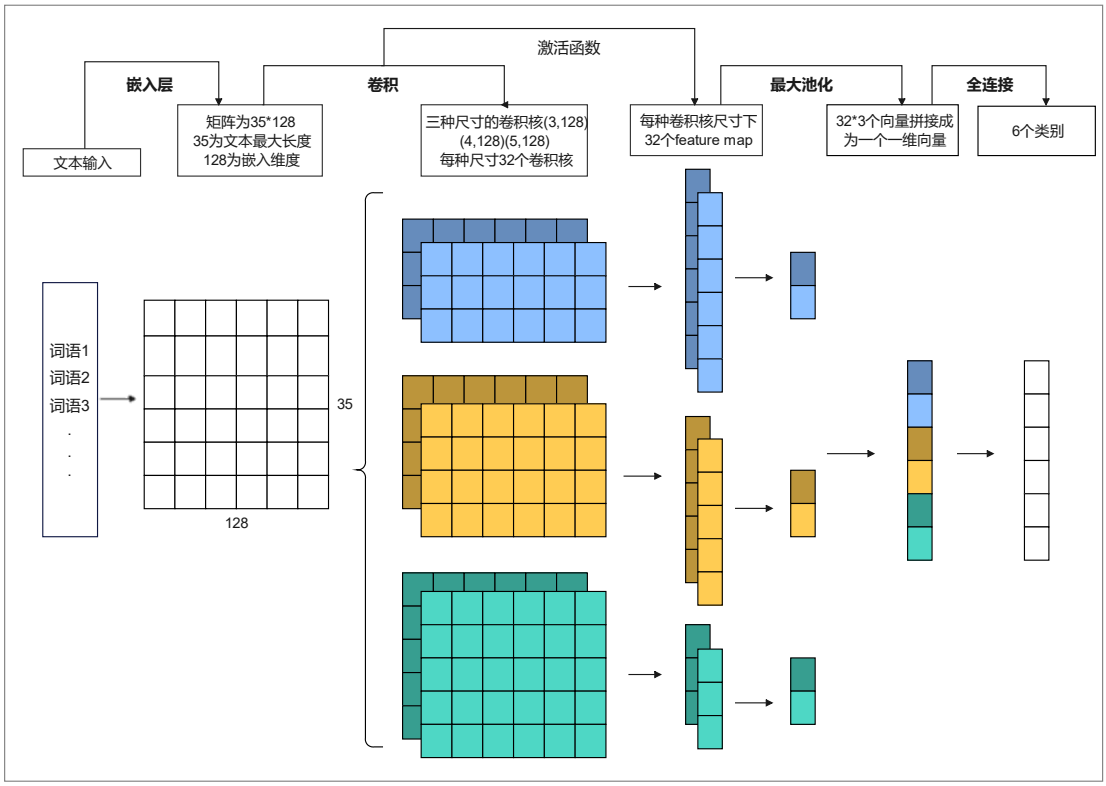

图 2 TextCNN 模型结构及参数设置

4.1.2 融合注意力机制和文本变长处理的 BiLSTM

BiLSTM 模型无法自主确定序列中的关键词，故引入注意力机制，使模型将注意力集中在需要重点关注的目标上，分配更多的权重，获取目标的更多细节信息，忽略不重要的信息。其计算公式为：

$$\varphi\left(h_t, h_s^i\right) = h_t^T W_a h_s^i \tag{7}$$

$$\alpha_i = \frac{\exp\left(\varphi\left(h_t, h_s^i\right)\right)}{\sum_{j=1}^{t} \exp\left(\varphi\left(h_t, h_s^j\right)\right)} \tag{8}$$

$$c = \sum_{i=1}^{t} \alpha_i h_s^i \tag{9}$$

其中，h_t表示最后一个时刻的隐藏层状态，表示针对全句提取到的特征向量。h_s^i表示i时刻的隐藏层状态，表示针对各个词语提取到的特征向量。通过式(7)计算两者之间的相关性得分，并如式(8)所示使用 softmax 函数将其转换为关注度权重，最后如式(9)加权平均得到关注关键信息后的全局特征向量。

在采用 BiLSTM 训练文本序列数据时，会遇到序列样本长短不一的情况，需要对样本序列进行填充，保证各个样本长度一致。但序列中添加很多无效值一方面浪费计算资源，另一方面也会影响特征提取。因此，在保证每个批次的数据按照文本长度降序排序的条件下，将序列输入模型前使用 pack_padded_sequence，将填充的无效值压缩，此时模型会根据该批次中不同时间步下的样本数量重新自动设置 mini-batch。在模型完成训练提取到特征向量后，为方便后续操作需要使用 pad_packed_sequence，把压紧的序列再填充回来。

综上，本文设计的 BiLSTM 模型结构如图 3 所示：

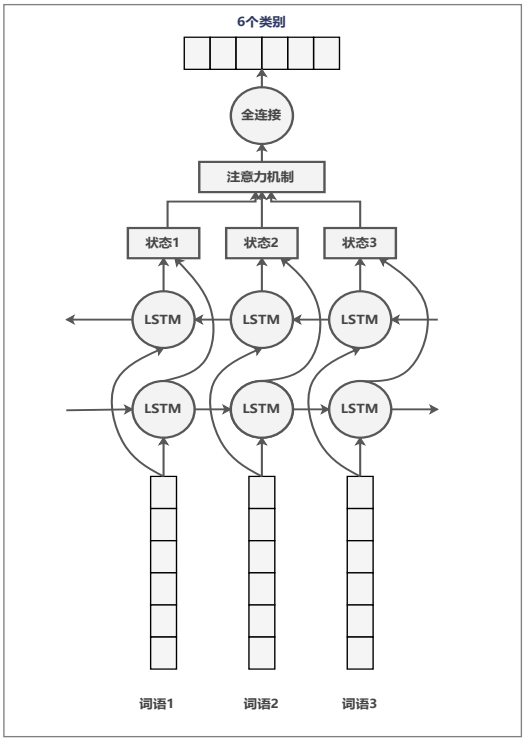

图 3 BiLSTM 模型结构

4.1.3 CNN-BiLSTM

相较于 BiLSTM-CNN，先进行全局特征的提取再进行局部特征提取的思路。CNN-BiLSTM 先利用 CNN 提取文本局部的特征，后使用 BiLSTM 理解句子的语义信息。先通过 CNN 对局部词汇进行特征提取，获得局部显著特征，然后通过 BiLSTM 实现特征的全局理解，同时引入注意力机制帮助确定关键信息，最后使用全连接完成分类。图 4 展示了本文研究建立的 CNN-BiLSTM 网络。

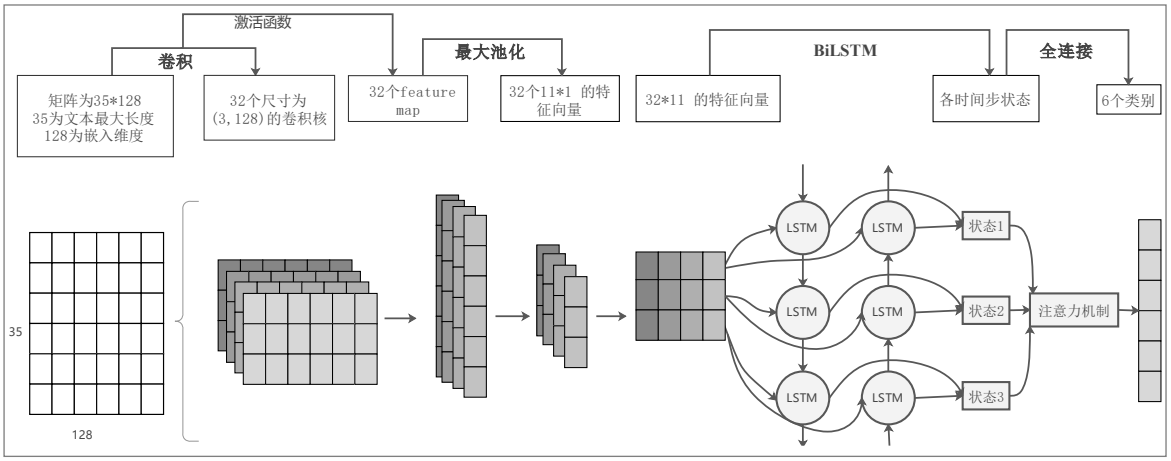

图 4 CNN-BiLSTM 结构及参数设置

4.2 模型评价指标

本文研究问题为多分类问题，指标选用宏观精确率(Micro Precision)、宏观召回率(Micro Recall)、宏观 F1 值(Micro F1)。评价时分别将各类别的评价转换为二分类问题评价，计算相关指标，最后用求平均的方式得出最终评价结果。

4.2.1 宏观精确率(Micro Precision)

精确率是指在预测为该类别的样本中真正属于该类别的样本所占的比例，在该问题中，实际意义为预测为某种情感的样本中真正属于该情感的样本的比例。计算时，首先分别计算每个类的精确度，再求平均值。具体计算过程如下。

$$Precision_i = \frac{TP_i}{TP_i + FP_i} \tag{10}$$

$$Micro\ Precision = \frac{\sum_{i=1}^{n} Precision_i}{n} \tag{11}$$

4.2.2 宏观召回率(Micro Recall)

召回率指在所有的该类样本中被预测为该类的比例，在该问题中，实际意义为预测为某种情感的样本中被预测为该情感样本的比例。计算时，首先分别计算每个类别的召回率，再求平均值。具体计算过程如下。

$$Recall_i = \frac{TP_i}{TP_i + FN_i} \tag{12}$$

$$Micro\ Recall = \frac{\sum_{i=1}^{n} Recall_i}{n} \tag{13}$$

4.2.3 宏观 F1 值(Micro F1 score)

因为 Precision 和 Recall 是一对相互矛盾的量，当 Precision 高时，Recall 往往相对较低，当 Recall 高时，Precision 往往相对较低，所以为了更好的评价分类器的性能，一般使用 F1-Score 作为评价标准来衡量分类器的综合性能。具体计算过程如下。

$$\text{Micro F1 score} = \frac{2 * Micro\ Precision * Micro\ Recall}{Micro\ Precision + Micro\ Recall} \tag{14}$$

4.3 实验结果

BiLSTM-CNN 模型参数设置如表 2 所示。该模型容易出现过拟合的问题，在参数设置时，尤其需要在 CNN 池化后加大 dropout 以提升其泛化性能。

表 2 模型参数设置

	参数名称	参数值
1	batch_size	128
2	embedding_dim	64
3	kernel size	3
4	filter_num	32
5	pool size	3
6	hidden_num	128
7	epoch	30

训练过程中，训练集和验证集的损失函数曲线和准确率曲线如图 5 图 6 所示，模型在训练集损失值收敛于 0.007 附近，验证集损失值有波动但未出现明显的过拟合现象。

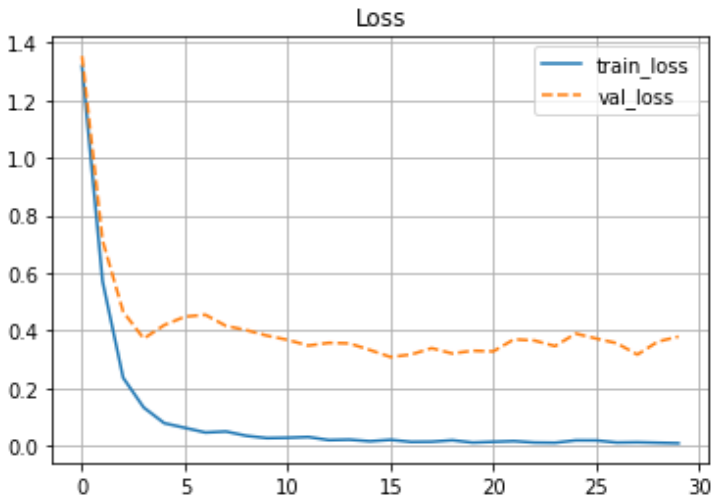

图 5 BiLSTM-CNN 损失函数曲线

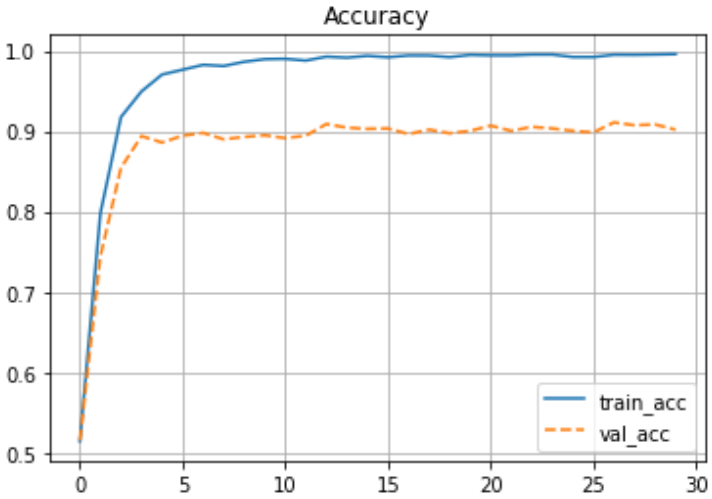

图 5 BiLSTM-CNN 准确率函数曲线

4.4 模型对比

表 3 为各模型在测试集上的分类效果。从测试集的分类结果看，BiLSTM-CNN 的精准率、召回率以及 F1 值都最高，融合注意力机制和文本变长处理的 BiLSTM 效果位列第二，TextCNN 和 BiLSTM-CNN 次之。

表 3 模型对比

模型	Micro Precision	Micro Recall	Micro F1 score
TextCNN	0.8417	0.8254	0.8329
BiLSTM	0.8681	0.8581	0.8618
CNN-BiLSTM	0.8266	0.8375	0.8310
BiLSTM-CNN	**0.8830**	**0.8782**	**0.8804**

结果表明，通过结合 BiLSTM 和 CNN 可以同时利用 CNN 识别局部特征与 BiLSTM 识别文本序列的能力，提升分类效果。但两者如何组合使用也十分关键，对比 BiLSTM-CNN,CNN-BiLSTM 分类效果最差，可能最初 CNN 层卷积池化时丢失了文本序列的一些信息，没有真正的提取到有效的特征，因此 BiLSTM 层的作用被大大减弱，导致最终实验效果较差。BiLSTM-CNN 模型和融合注意力机制和文本变长处理的 BiLSTM 模型分类效果都较好。因此，BiLSTM 层能更加有效地提取和保留整个文本的特征，该层的输出不仅包括当前词语特征，也包括上下文的特征，特征向量全面而有效。之后加入 CNN 层使得局部特征被加以提取放大，实现了最优分类效果。

5 总结

本文针对文本情感多分类问题展开研究，对比多种模型发现 CNN 和 BiLSTM 模型结合使用能够同时兼顾局部特征和全局特征。先使用 BiLSTM 提取全局特征，后使用 CNN 提取局部特征的 BiLSTM-CNN 分类效果最优。但仍然会出现过拟合以及训练后期验证集损失函数上升的情况。本研究考虑从以下三个方面进

行改进：首先，在文本数据向量化表示时，不简单地使用 embedding 完成，尝试使用不同的词向量转换方式，如 Word2vec、BERT 以及 Glove 等。其次，针对 BiLSTM-CNN 中的卷积操作进行改进，减少过拟合风险。最后，目前该模型针对 BiLSTM 和 CNN 的融合是采取串联模式，可以尝试采用并行模式，两个模型分别位于两个独立的通道，各自计算得出特征向量后，采用注意力机制等方式进行融合，这样可以避免出现因前一个模型特征提取低效对后一个模型的影响。

6 参考文献

[1]孙炜. 基于 Attention-CNN 的文本情感分类研究[D].西安科技大学,2021.
[2]余静莹. 基于双向长短期记忆网络的文本情感分类算法研究[D].安徽大学,2021.
[3]张索宇. 基于注意力机制和神经网络融合的文本情感分类[D].重庆理工大学,2021.
[4]赵昌健. 基于注意力机制的文本情感分类研究[D].电子科技大学,2021.