自然语言处理

姓 名：周子凯

学 号：2201934

专 业：计算机科学与技术

2023 年 1 月 6 日

一 第三方库配置

tensorflow-1.14.0

keras-2.6.0

genism-4.1.2

二 实验内容

首先数据集来源于网络资源，由 datasense[1]个人博客博主整理完成，包含 10000 条均衡的正负情感评论数据，数据来源于电子商城服装区的评论与售后反馈。

第一阶段：基于给定数据集，首先尝试使用基于多层 LSTM 的网络模型解决情感二分类问题，在词嵌入方法上选择最简单与普遍的词袋模型,基于词袋模型，尝试 n-gram（n=1，2，3），unigram 在词袋模型基础上使用 SVM，逻辑回归，随机森林等机器学习算法组合 Grid Search 寻优，具体来说，在逻辑回归模型上使用 L1 正则化或 L2 正则化，在 SVM 模型上对惩罚因子与 gamma 进行区间寻优。

第二阶段：经过对相关文献与研究的了解，引入 Word2ec 模型用于替代词袋模型，引入 Bidirectional LSTM 替换 LSTM 在以上提及的所有组合中在交叉验证时，Word2Vec-BILSTM 模型测试集准确率最高，最高达 95.23%。

第三阶段：经过对基于规则的情感分类研究的了解，从一篇基于否定语和情感词典提升机器学习分类性能的文献中引入否定语义的概念，在经过合适的处理后，情感词典组合否定语义达到了 90%以上的准确率，说明数据质量优良，且该方法可行。

第四阶段：出于集二者之长的想法，从相关技术论坛上了解到多输入并行的网络结构提升了对模型的使用空间，在否定语义与情感词的基础上整合情感分数与语义分数，以及情感词个数等作为基于规则获得的语义信息；在 Word2Vec 的基础上获得词向量信息，通过构建独立并行的模型结构接收两种输入，整理出一种输出。

第五阶段：基于第四阶段的两种输入信息，分别设计适合的并行模型，在模型设计时，引入注意力机制，结合注意力机制与堆叠 BI-LSTM 处理词向量信息，使用单层 LSTM 来处理语义信息（原始就是数字特征）

第六阶段：在原模型的基础上，调整分类阈值，学习率，学习率衰减策略，批次大小，句子最大长度等，在 BILSTM 模型参数确定时，在堆叠 BILSTM 上使用基于遗传优化的改

进算法确定了层数以及每层的参数（具体数字近似至 2 的 n 次方，效果几乎无差别），最终使用早停的方式获得测试集上的最佳模型。

三 模型比较

实验中使用到的分类器准确率总结如下表 5-5 所示：

表 0-1 分类器性能比较

Table 0-1 Classifier performance comparison

分类器	Accuracy
SVM	0.831
Random Forest	0.879
情感词典+否定语义	0.925
Logistic Regression	0.933
Logistic Regression after GS	0.939
Word2Vec-LSTM	0.948
Word2Vec-BI-LSTM	0.952
Word2Vec-LSTM+BI-LSTM 自定义模型	0.970

在尝试机器学习算法时，Random Forest 与 SVM 的训练效果并不显著。在维度高的大样本集上此现象是可以理解的。本节只展示预测准确率可超过 90%的部分模型。

1）情感词典

在二分类理论中，分类阈值（分类的标准，阈值的作用简单理解：比阈值小，类别为 0；比阈值大类别为 1）。在使用情感词典时，可调整情感词典分类阈值为 2 划分正负标签，正确率可达 0.925；从图 6-3 与图 6-4 可以看出，否定语义对情感词典起到有效的辅助作用，可以将精度提升约 7%。

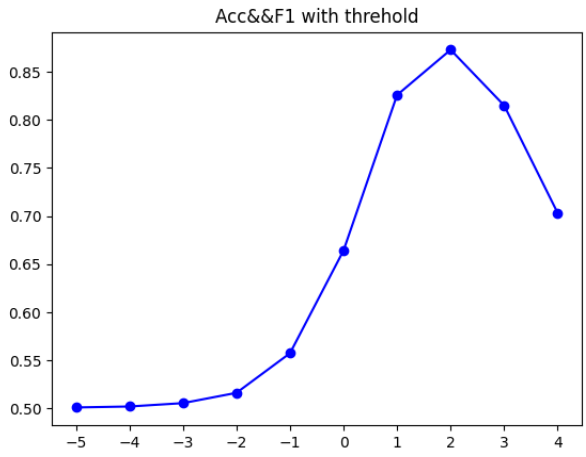


图 3-1 情感词典（随分类阈值变化）

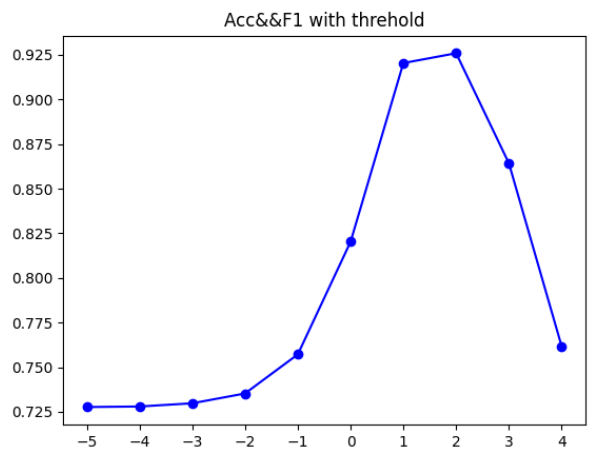


图 3-2 情感词典+否定语义（随分类阈值变化）

2）逻辑回归

从混淆矩阵中可以看出有 36 个 FN，28 个 FP，测试数据集总大小为 1000 条。此模型准确率约 93%；学习率曲线结果说明在数据集接近 8000 时，模型得分均值仍可在 90%以上（得分不等于准确率）

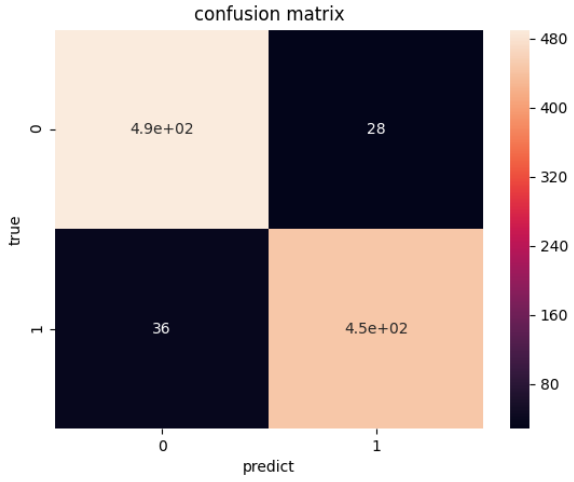


图 3-4 逻辑回归混淆矩阵

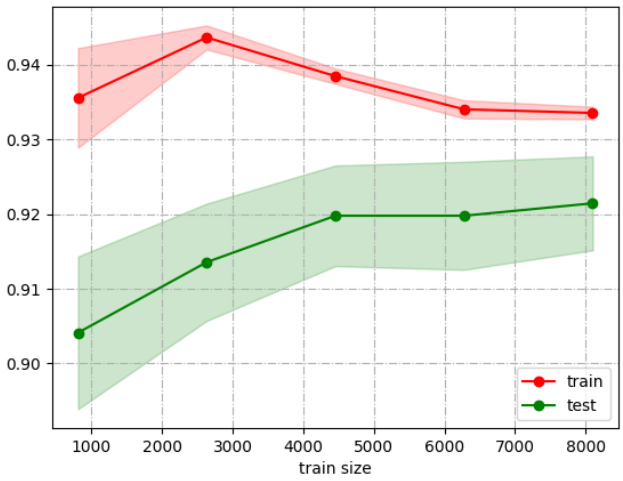


图 3-1 逻辑回归学习率曲线

从下面的 ROC 图像上看，左下方面积为 0.97，0 这个指标是相对而言的，0.97 很接近 1，说明模型性能良好。

同理从 PRC 图像上看，右下方面积为 0.97，如 ROC 图像，作为参考。

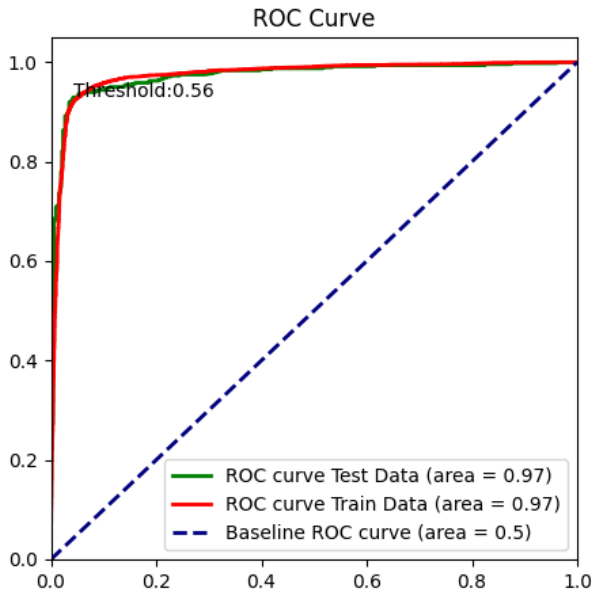


图 3-6 逻辑回归 ROC 曲线

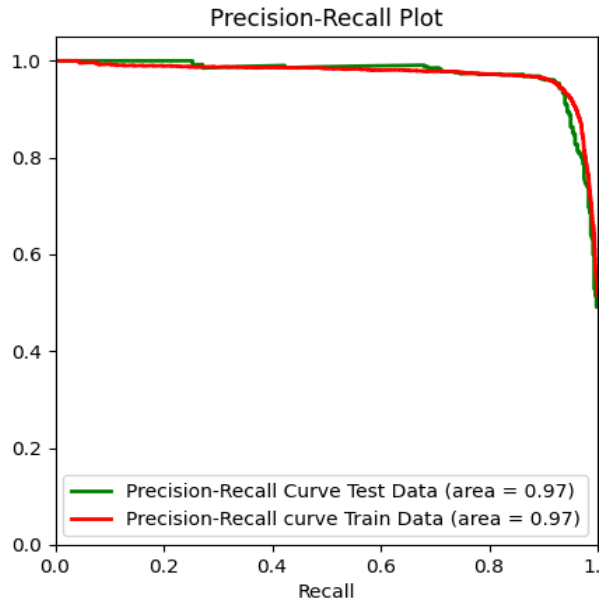


图 3-7 L2 逻辑回归 PRC 曲线

3）基于 GS 优化的逻辑回归

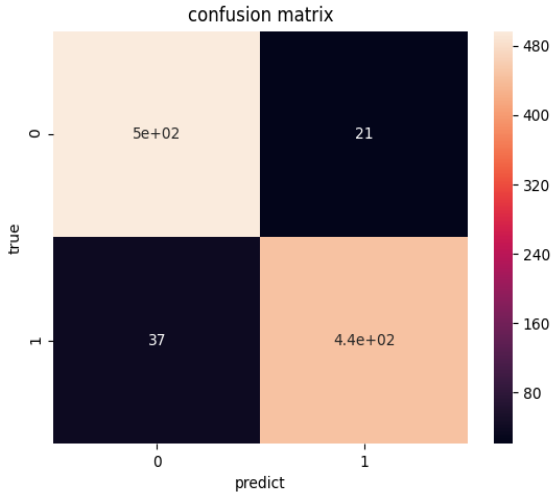


图 3-8 优化后 LR 混淆矩阵

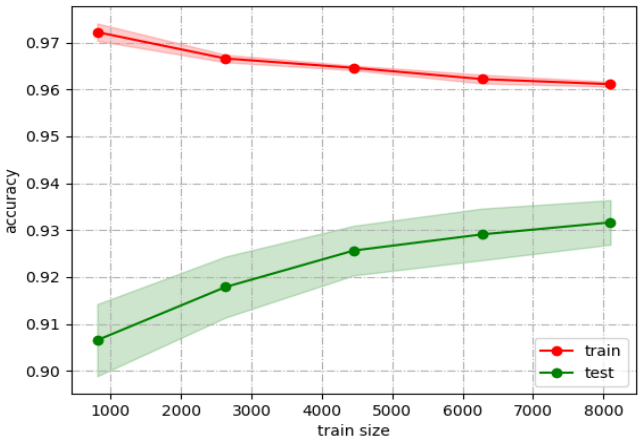


图 3-2 优化后 LR 学习率曲线

从混淆矩阵中可以看出有 36 个 FN，28 个 FP，测试数据集总大小为 1000 条。此模型准确率约 94%，正则化对逻辑回归模型是有效的；学习率曲线结果说明在数据集接近 8000 时，模型得分比逻辑回归模型要好。

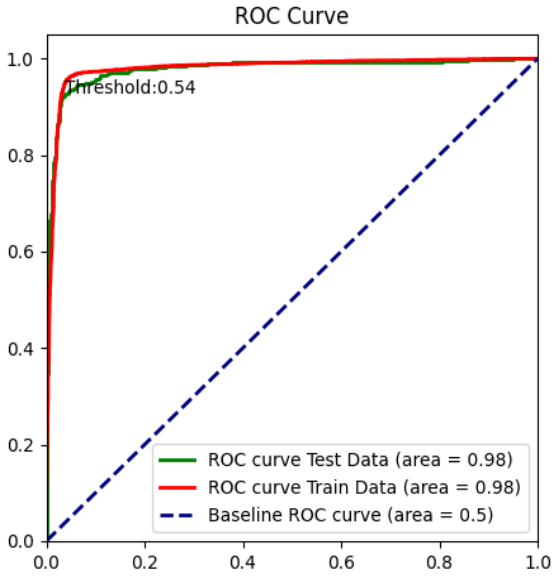


图 3-10 优化后 LR ROC 曲线

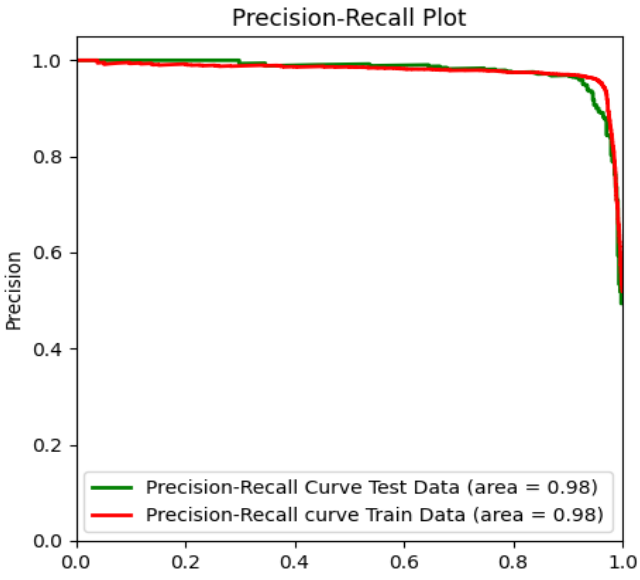


图 3-11 优化后 LR PRC 曲线

从上面的 ROC 曲线与 PRC 曲线上看，面积值大小为 0.98，比 0.97 更大，更接近于 1。说明优化后逻辑回归模型较逻辑回归模型分类性能更佳。

4）WordVec-LSTM 模型

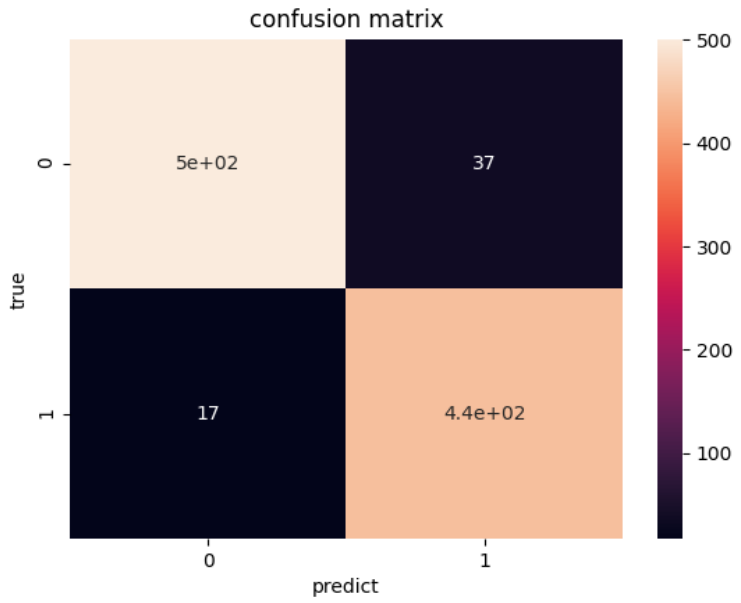


图 3-12 Word2Vec-LSTM 模型混淆矩阵

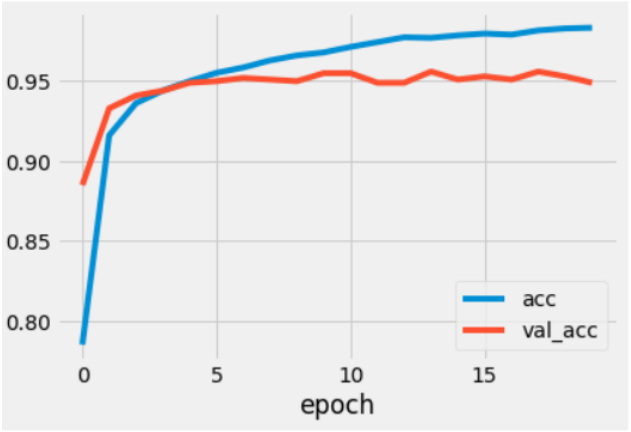


图 3-13 Word2Vec-LSTM 模型准确率变化曲线

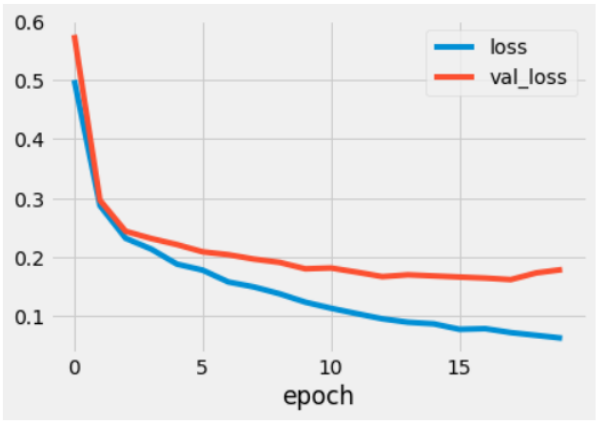


图 3-14 Word2Vec-LSTM 模型损失变化曲线

从混淆矩阵中可以看出，有 37 个 FP，17 个 FN，测试集数据大小为 1000 条。此模型准确率约 94.8%.；从准确率变化曲线中可以看出在 20 次训练之后，预测集准确率与测试集准确率都可以达到 95%附近；从损失变化曲线中可以看出，模型损失整体降低，但是并不是一直降低。但是 Word2Vec-LSTM 模型性能还是可以保证大于 90%，接近 95%的。

5）Word2Vec-BI-LSTM 模型

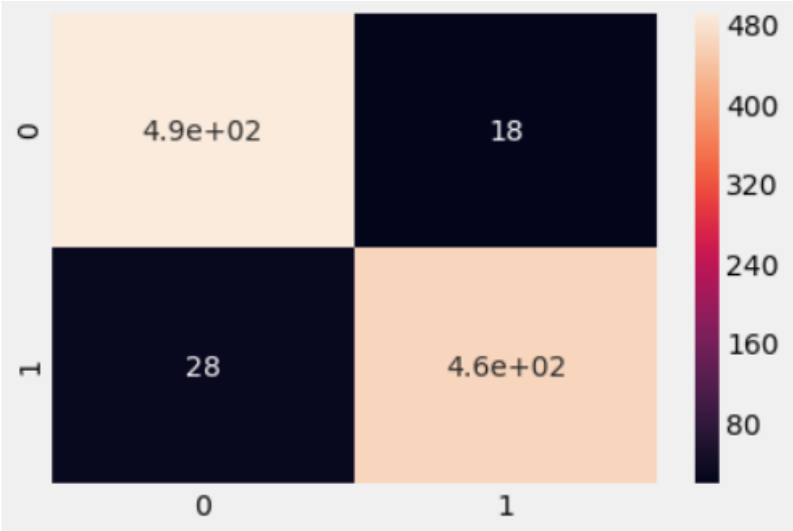


图 3-15 Word2Vec-BI-LSTM 模型混淆矩阵

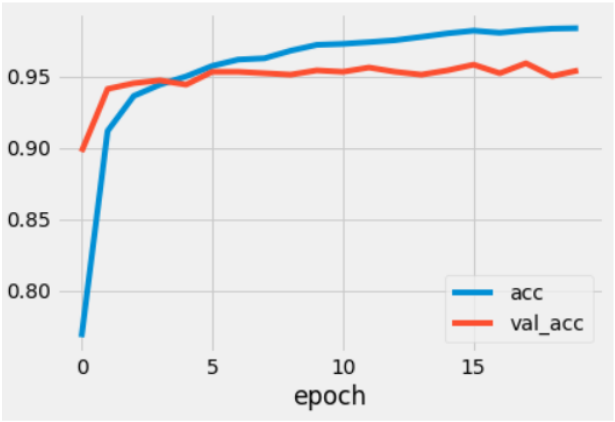


图 3-16 Word2Vec-B-LSTM 模型准确率曲线

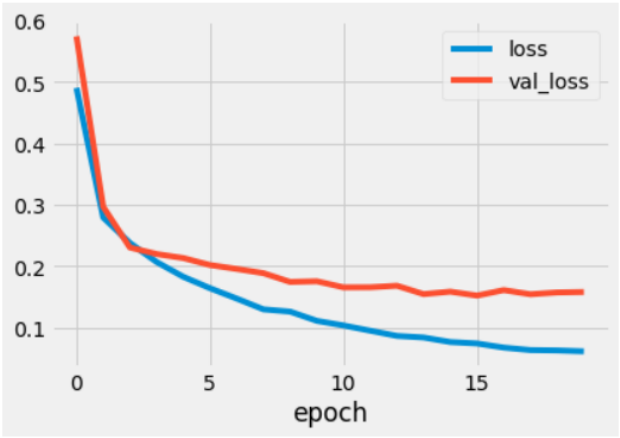


图 3-17 Word2Vec-LSTM 模型损失变化曲线

从混淆矩阵中可以看出，有 18 个 FP，28 个 FN，测试集数据大小为 1000 条。此模型准确率约 95.2%.；从准确率变化曲线中可以看出在 20 次训练之后，预测集准确率与测试集准确率都可以达到 95%附近；从损失变化曲线中可以看出，模型损失整体降低，较 Word2Vec-LSTM 而言，有稍微提升，但是变化不明显，说明 Word2Vec 结合传统 LSTM，BI-LSTM 构建的神经网络模型的性能基本上接近 95%，在 95%左右徘徊。

8）提出的并行模型

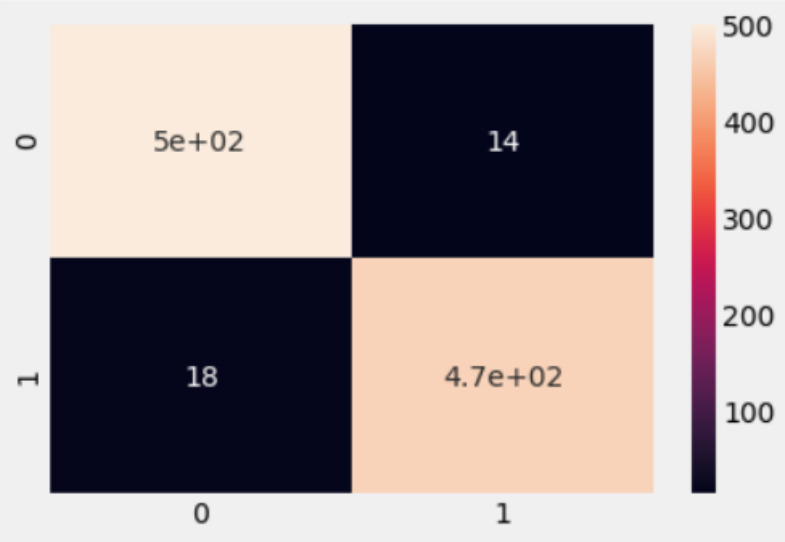
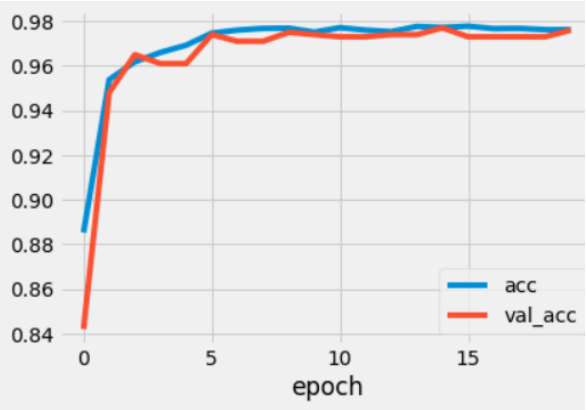


图 3-18 提出的模型混淆矩阵



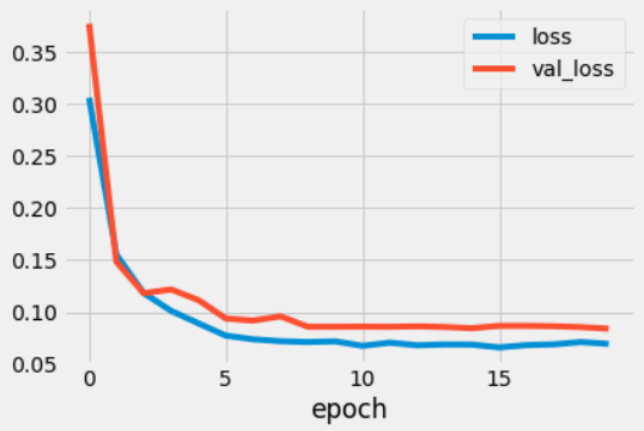


图 3-19 自定义模型准确率变化曲线

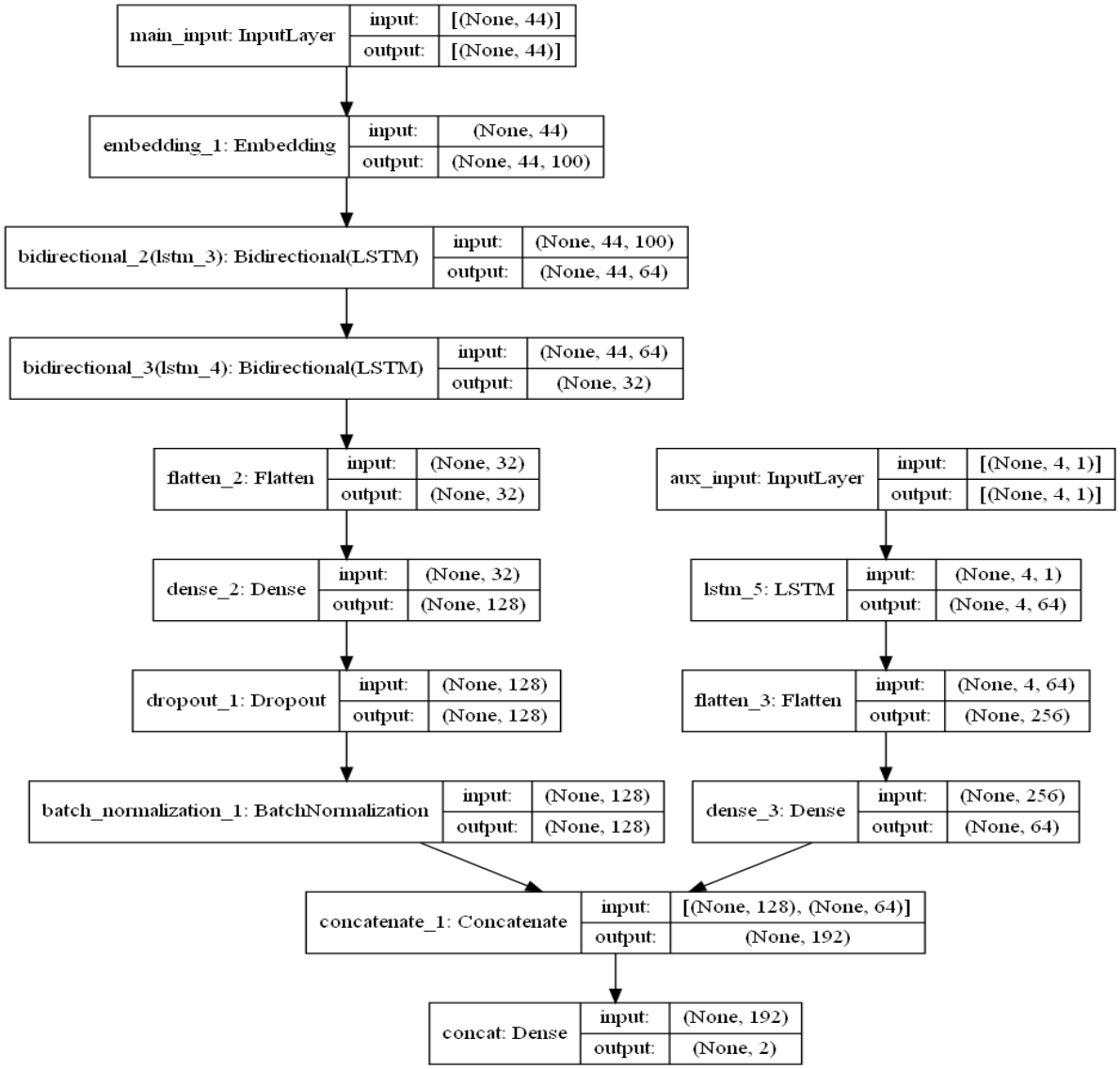
Figure 0-1 User-defined Model accuracy curve

图 3-20 自定义模型损失变化曲线

Figure 0-2 User-defined Model loss curve

自定义模型结合 LSTM 与 BI-LSTM，结合情感信息向量与 Word2Vec 词向量的信息，从混淆矩阵中可以看出，有 14 个 FP，18 个 FN，测试集数据大小为 1000 条。此模型准确率约 97%.；从准确率变化曲线中可以看出在 20 次训练之后，预测集准确率与测试集准确率都可以达到 98%附近；从损失变化曲线中可以看出，模型的损失一直降低。无论从性能，还是稳定性上自定义的模型都是最佳的选择。

融合模型结构示意图



数据集来源：https://www.mlln.cn