**警情记录文本分类**

**个人信息**

**2023春-计算机应用技术 231017000159 毕莹**

**一、业务目标**

1、提升警情记录处理效率：通过自动分类警情记录，可以减少人工分类的时间和精力，提高警情记录处理的效率。

2、提高警情分析准确性：深度学习模型能够学习并理解警情记录中的文本信息，从而更准确地分类和标记警情，为警情分析提供更准确的数据支持。

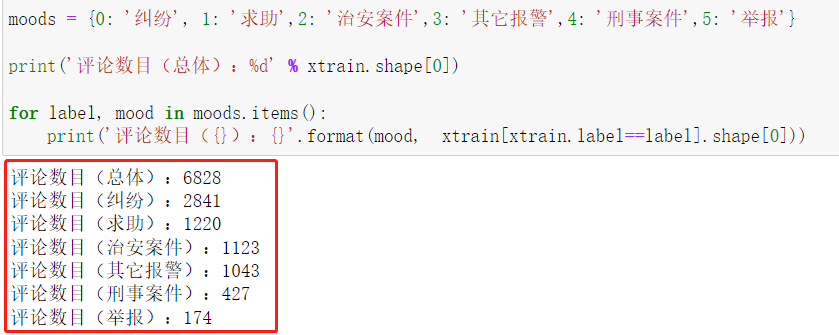
3、促进警务智能化发展：基于深度学习的警情记录文本分类技术是警务智能化发展的重要方向之一，可以为警务决策提供更准确、快速的支持。

**二、数据处理流程**

1、准备数据：出警情况共分为5类，每一类对应一个数字标签值，分别为：

'纠纷':0, '求助':1, '治安案件':2, '其它报警':3,'刑事案件':4,'举报':5

人工标注与审核各报警内容

统计各警情类别的分布比例，如下图所示：

2、数据清洗与过滤：剔除英文、数字，以及空格；剔除英文标点符号和特殊符号；剔除中文标点符号

3、分词后去除停用词：将文本警情内容分词，用于构建词频字典，同时去除停用词等无意义词项

5、将词频字典写入文件中保存

样本在词频字典中的向量矩阵

1. 模型定义：使用tensorflow开源框架

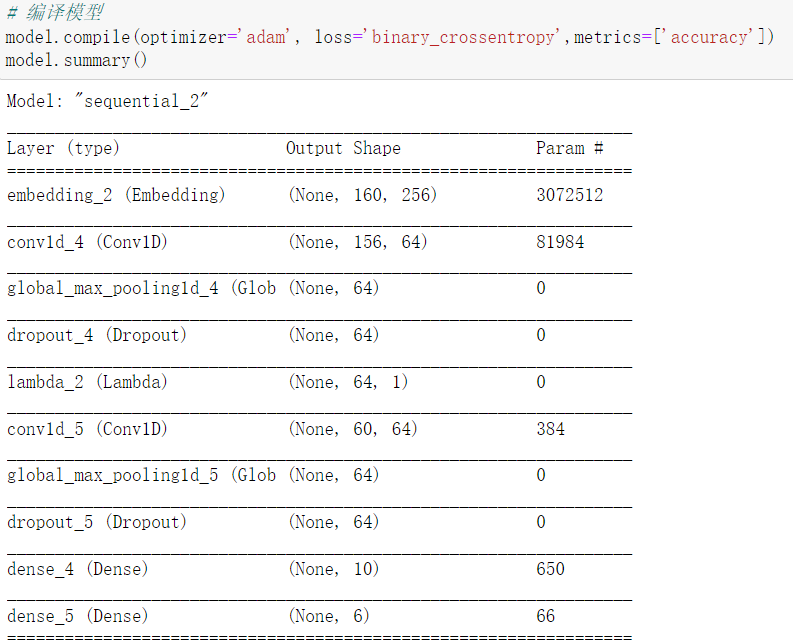
7、模型训练与调整参数

8、测试集测试结果

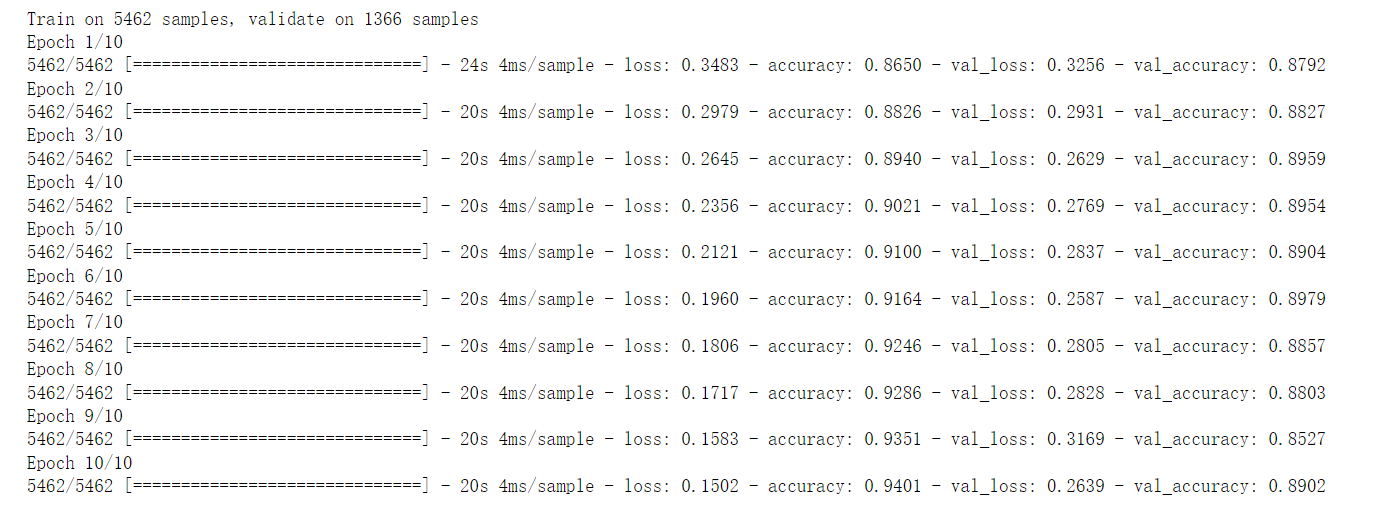
**三、模型描述**

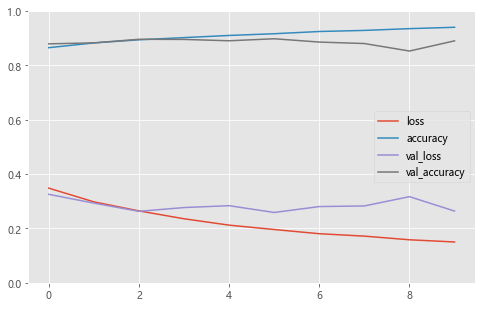
变量定义和模型构建

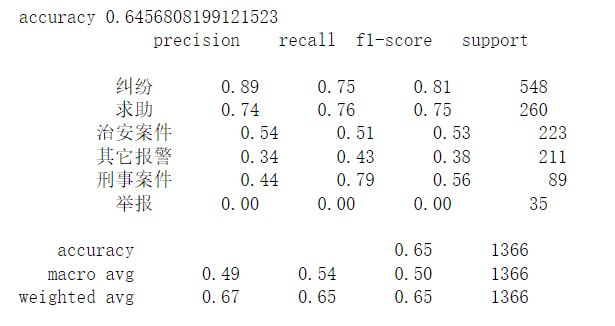




**四、最终结果**







数据表格显示了几个分类的准确性（accuracy）、精确度（precision）、召回率（recall）和 F1 分数，以及每个类别的支持度（support）。

**accuracy：**这是模型在所有样本上的总体准确率。模型的总体准确率是 0.65。这意味着模型正确分类了大约 65% 的样本。

**precision：**模型在预测为某一特定类别时，实际属于该类别的样本所占的比例。对于“纠纷”类别，模型的精确度是 0.89，这意味着模型预测为“纠纷”的样本中，实际属于“纠纷”类别的样本占了 89%。

**recall：**模型在所有实际属于某一类别的样本中，正确预测为该类别的样本所占的比例。对于“求助”类别，模型的召回率是 0.76，这意味着在所有实际属于“求助”类别的样本中，模型正确预测为“求助”的样本占了 76%。

**F1-score：**精确度和召回率的调和平均值，用于综合评估模型的性能。F1 分数越高，模型性能越好。对于“纠纷”类别，模型的 F1 分数是 0.81。

**support：**某一类别的样本数量。“纠纷”类别的样本数量是 548，“求助”类别的样本数量是 260。

**macro avg：**每个类别的准确率、精确度和召回率的简单平均值。macro avg 的目的是提供一个整体的性能评估，不考虑类别不平衡。macro avg 的准确率是 0.49，精确度是 0.54，召回率是 0.50。

**weighted avg：**每个类别的准确率、精确度和召回率的加权平均值，权重为每个类别的支持度。weighted avg 考虑了类别不平衡。在这个案例中，weighted avg 的准确率是 0.67，精确度是 0.65，召回率是 0.65。

从验证结果来看，模型在某些类别（如“纠纷”和“求助”）上的性能较好，但在其他类别上的性能还有待提升。

**小结**

本人大学毕业后虽然还从事计算机领域相关工作，但用的已经不太深入，之前学习过的内容只还记得一部分，通过这次课程的学习和作业的完成，也回顾了知识，同时也请教了老师，有些地方做的不好还请老师指正，谢谢。