**PYTHON大作业——虚假新闻检测**

**一、问题描述**

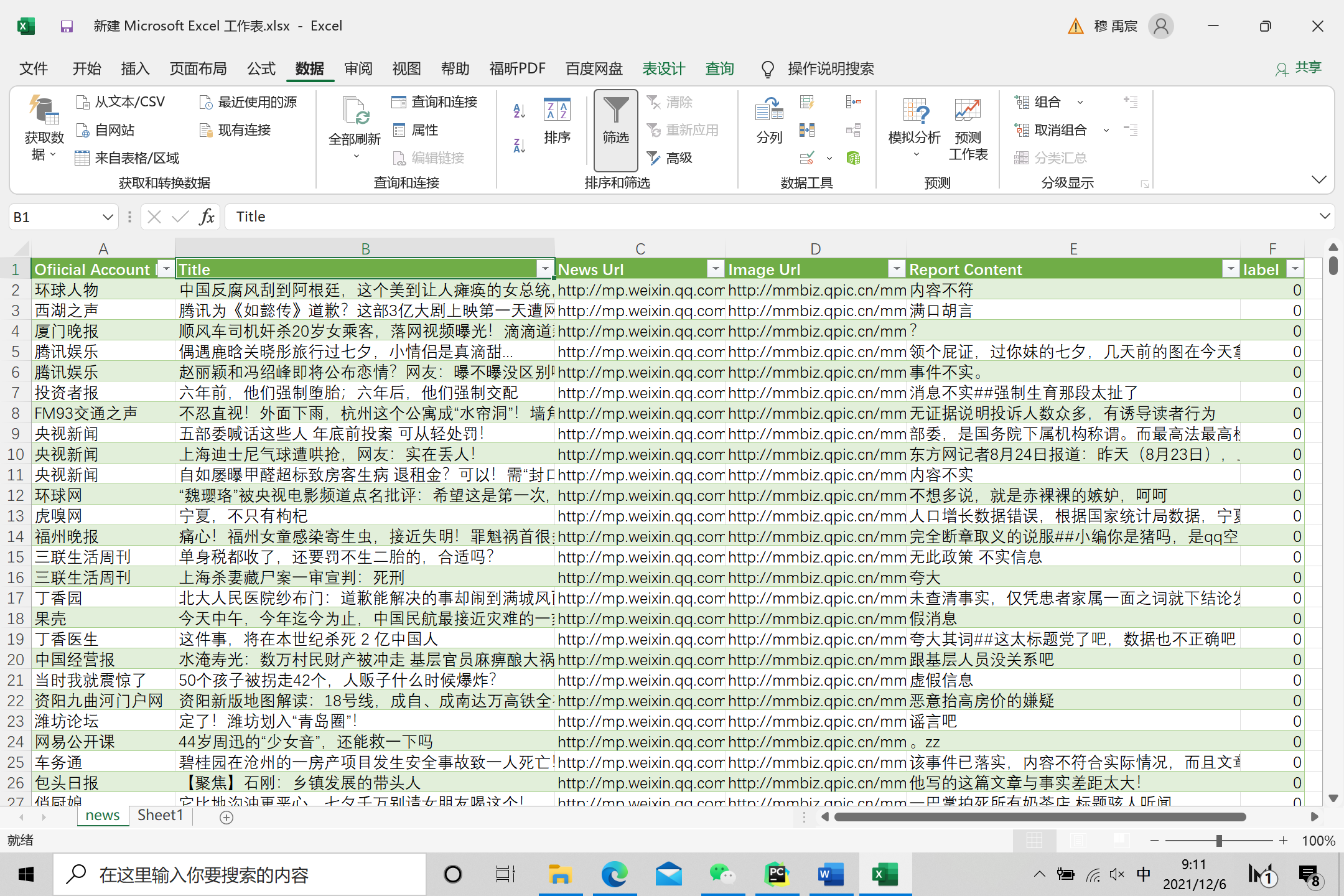
自然语言处理中的虚假新闻检测问题，本质上是一个文本二分类问题

定义：给定一个信息的标题、出处、相关链接以及相关评论，尝试判别信息真伪。

输入：信息来源、标题、相关超链接、评论

输出：真伪标签（0: 消息为真，1: 消息为假）

**二、数据集说明**



共6个标签，Official Account Name, Title, News Url, Image Url, Report Content, label

其中，Label0表示真新闻，1表示假新闻。

**三、方法介绍**

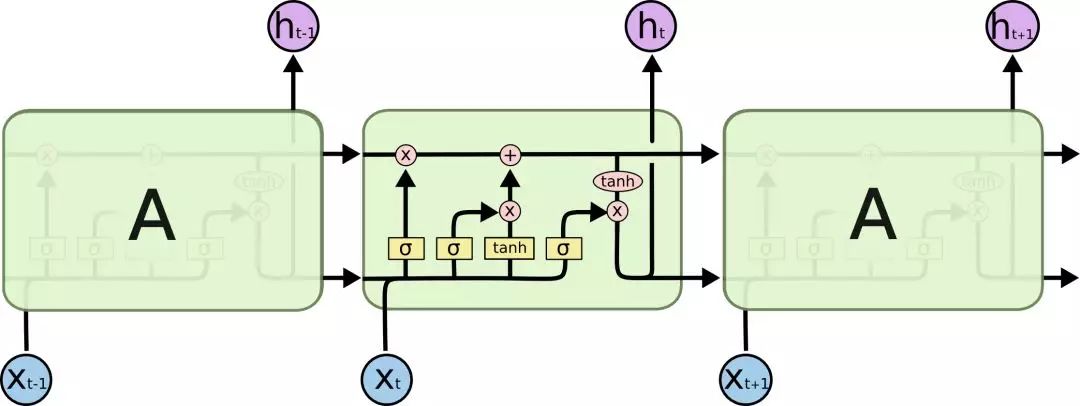
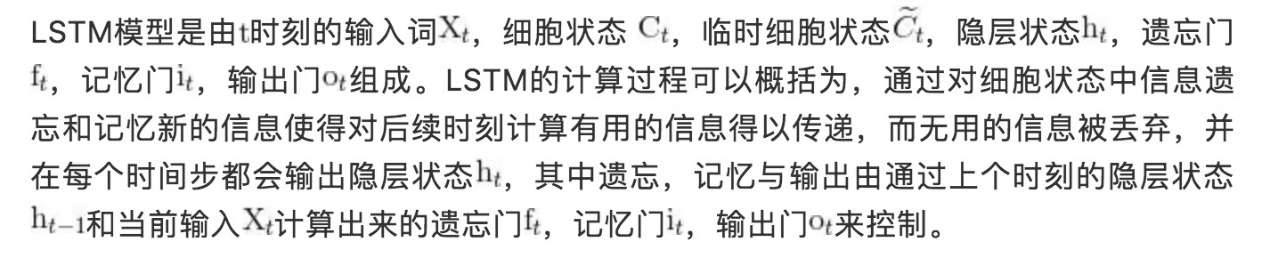
**深度学习方法**

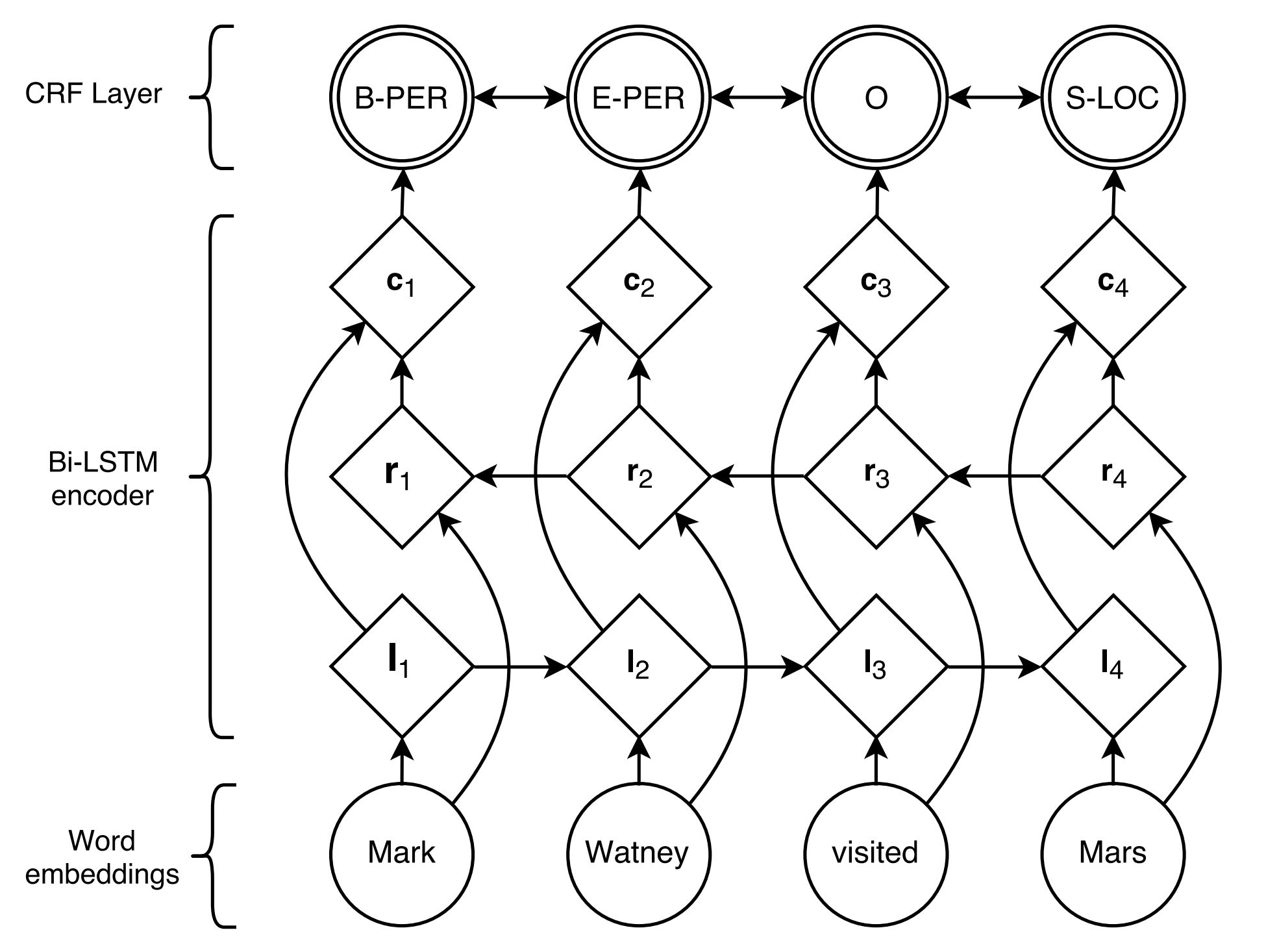
1.词嵌入

词嵌入实际上是一类技术，单个词在预定义的向量空间中被表示为实数向量，每个单词都映射到一个向量。将文本X{x12345……xn12345……yn }，这个映射的过程就叫做词嵌入。之所以希望把每个单词都变成一个向量，目的还是为了方便计算，通过词嵌入这种方式将单词转变为词向量，机器便可对单词进行计算，通过计算不同词向量之间夹角余弦值cosine而得出单词之间的相似性。

2.BILSTM神经网络

LSTM的全称是Long Short-Term Memory，它是RNN（Recurrent Neural Network）的一种。LSTM由于其设计的特点，非常适合用于对时序数据的建模，如文本数据。BiLSTM是Bi-directional Long Short-Term Memory的缩写，是由前向LSTM与后向LSTM组合而成。两者在自然语言处理任务中都常被用来建模上下文信息。

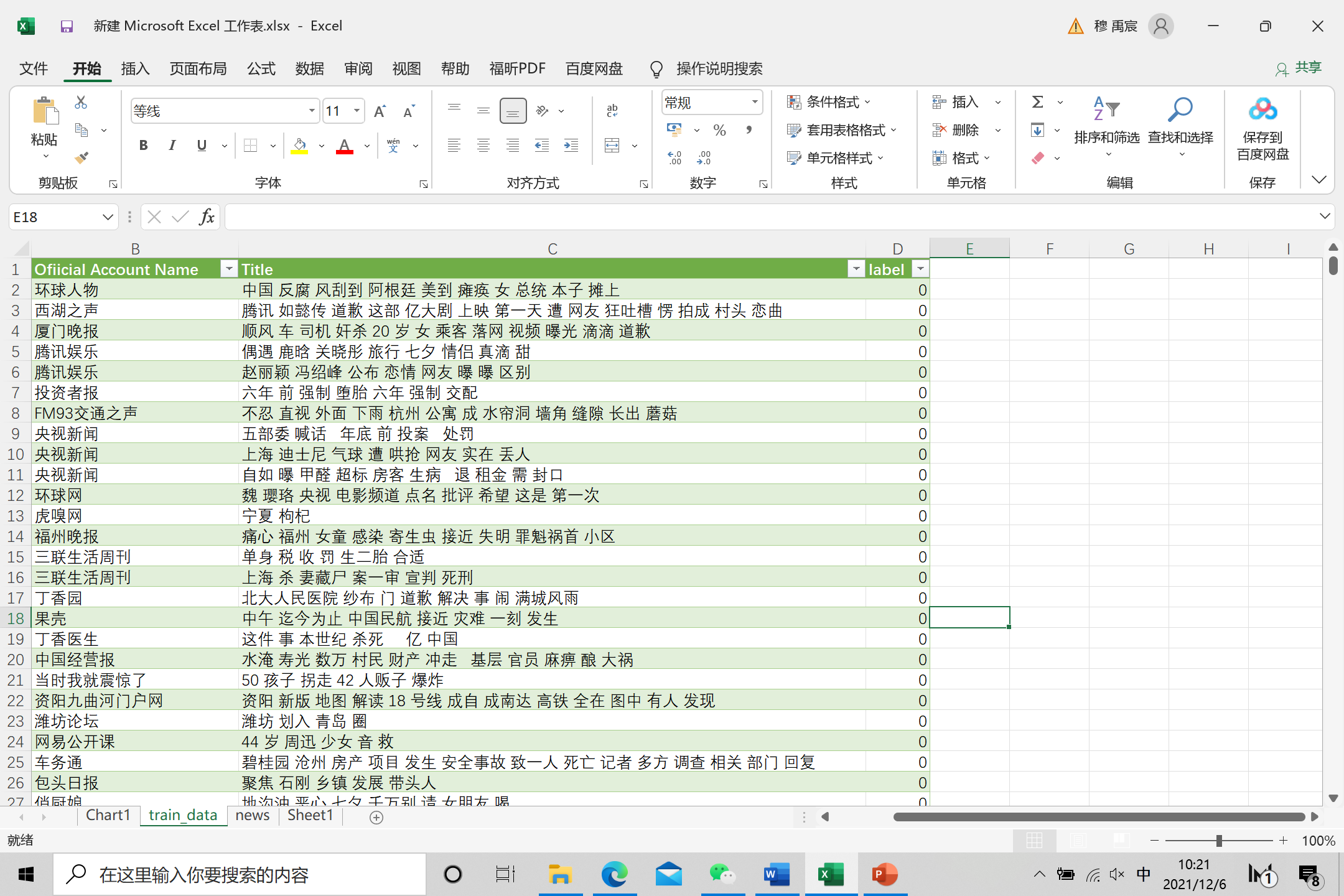


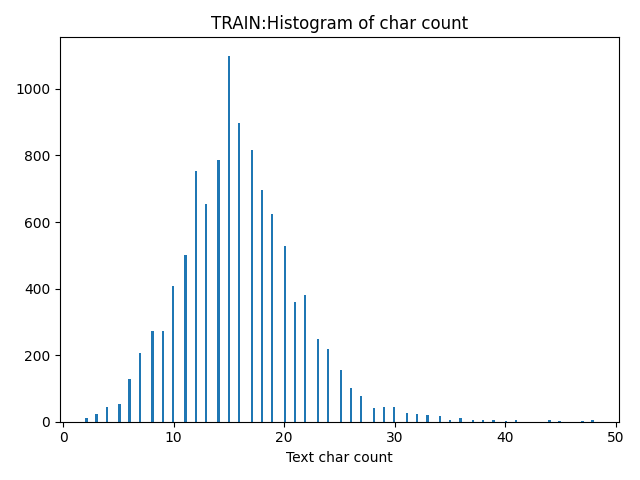


**四、关键代码细节**

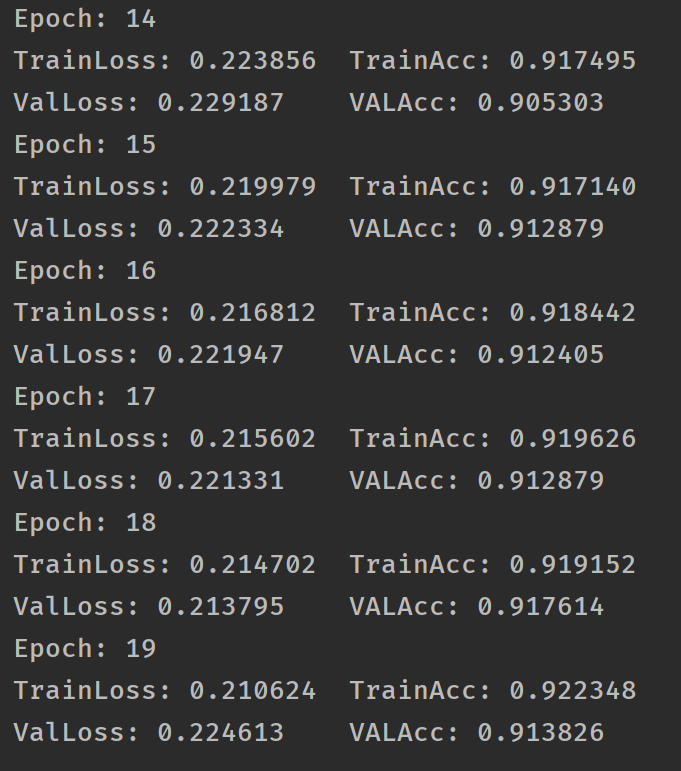
**五、运行截图**

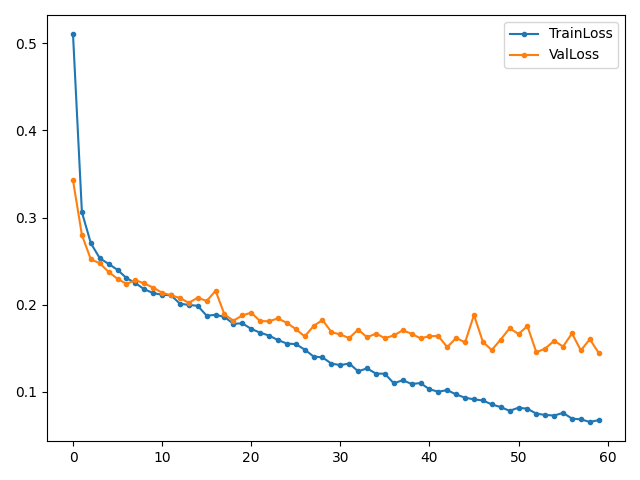
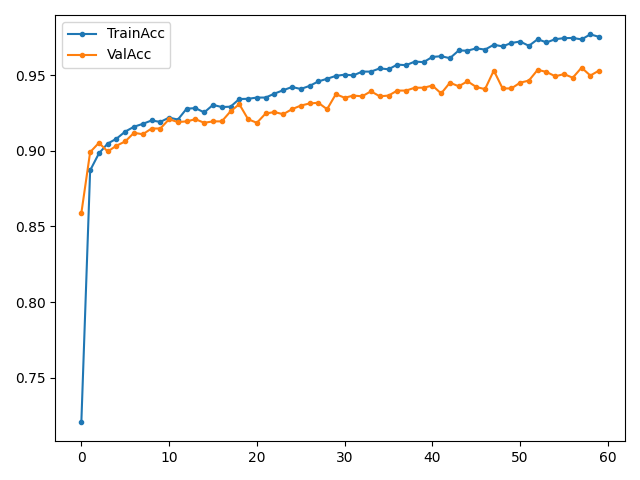
**1.经过预处理后的数据**



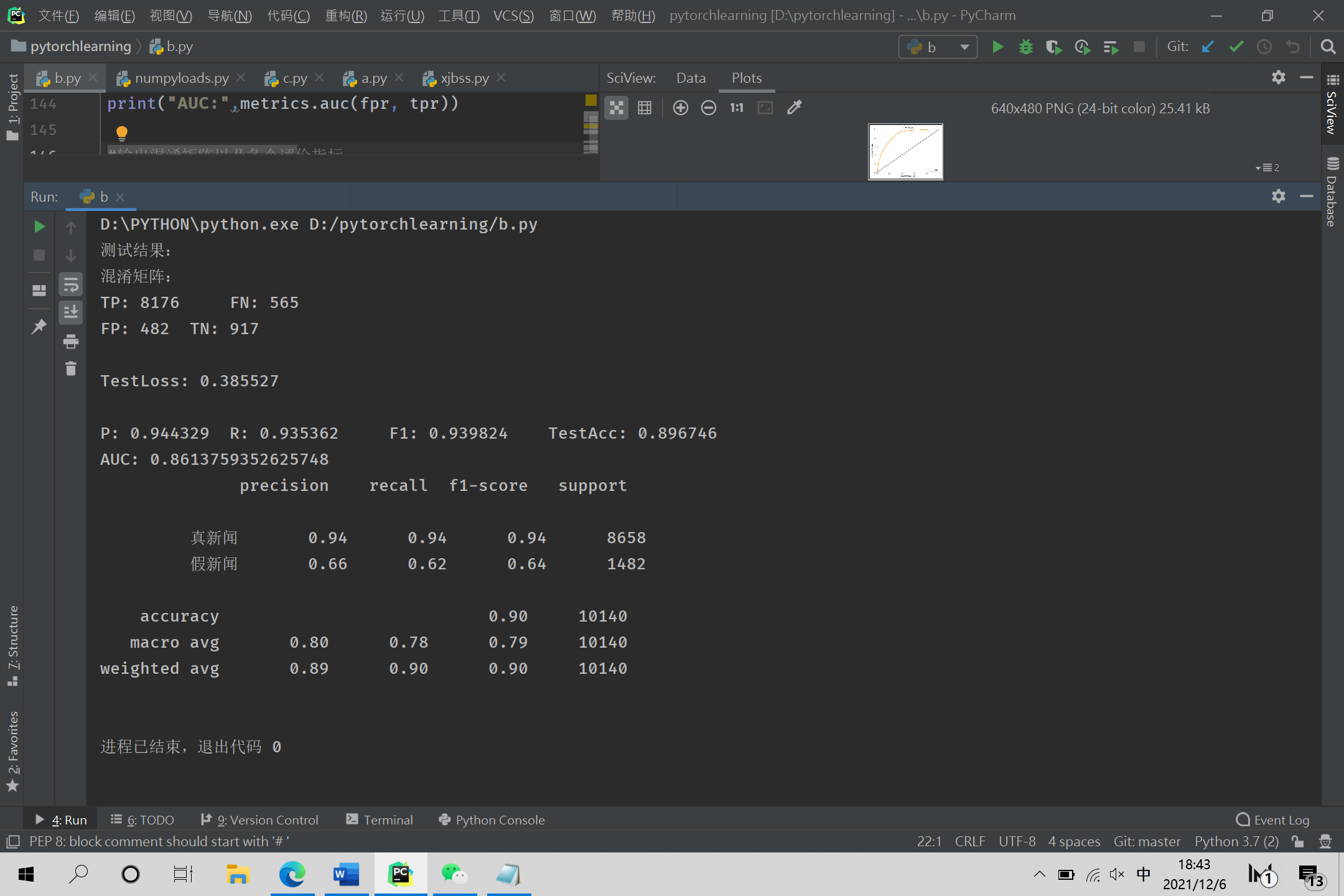


**2.训练神经网络过程中的部分结果**

****



**六、实验结果**

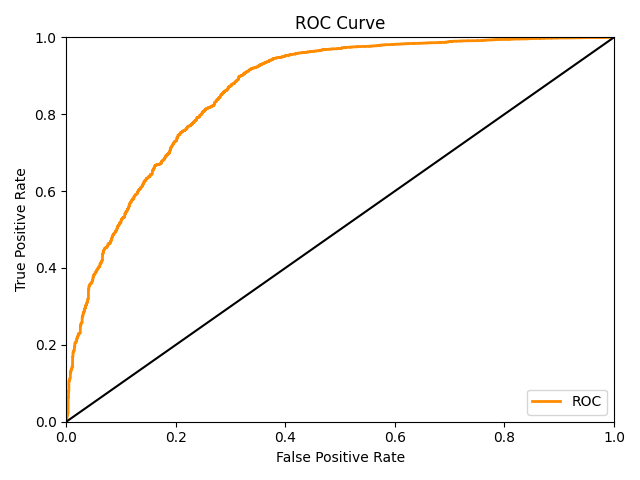


|  |  |  |
| --- | --- | --- |
|  | **PP** | **PN** |
| **P** | **TP=8176** | **FN=565** |
| **N** | **FP=482** | **TN=917** |

**Precision=94.4% Recall=93.5% F1=93.9%**

**Accuracy=89.6% AUC=86.1%**

**下图是ROC曲线**

****

**七、结果分析**

首先Accuracy=89.6%，证明我们预测对了89.5%的数据，这个数据是比较理想的，证明达到了训练的效果。

Precision=94.4%，真正为正类的占所有预测为正类的94.2%，说明我们将一少部分错误的预测为正确的，但数据仍然非常高

Recall=93.5%，表示有多少正类样本被预测正确，达到93.5%，也是很理想的数据

F1=93.9%，是P和R两者的调和平均，达到93.9%，也是很高的数值

AUC=86.1%，表示预测正例排在负例前的概率，达到86.2%，也算非常理想，从ROC曲线也能看出，比较靠近（0，1）这个点了。那么，这个分类器是很成功的，尤其对于我们的训练集非常不平衡的时候，这么高的AUC显然是非常好的数值。