**Intermediate task**

一、数据

数据集是中文微信消息，包括微信消息的Official Account Name，Title，News Url，Image Url，Report Content，label。Title是微信消息的标题，label是消息的真假标签（0是real消息，1是fake消息）。训练数据保存在train.news.csv，测试数据保存在test.news.csv。

二、任务(**建议使用Scikit Learn和PyTorch实现**)

1.(数据统计)分别统计训练集和测试集中真假样本的数量，填写下表1。

Table 1 . statistics of data

|  |  |  |  |
| --- | --- | --- | --- |
|  | #real | #fake | #total |
| train.news.csv |  |  |  |
| test.news.csv |  |  |  |

2.(验证集上调节参数)只根据微信消息的title文字，预测消息的真假。从train.news.csv中随机抽取2000个样本作为**验证集(validation set)**，剩余的数据作为**训练集(training set)**，使用test.news.csv中的样本作为**测试集(testing set)**。使用中文预训练词向量语料库[[1]](#footnote-1)把中文微信消息的title表示成词向量。根据训练集数据，训练**两层**的RNN, LSTM和GRU模型。模型隐层节点数量参数值的调节范围是{64,128,256}，通过在验证集上评价，得到的结果填入表2。

Table 2. results of intermediate tasks on validation set

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Model | Hidden | Label | Precision | Recall | F1 | Accuracy | AUC |
| RNN | 64 | Real |  |  |  |  |  |
| Fake |  |  |  |
| LSTM | 64 | Real |  |  |  |  |  |
| Fake |  |  |  |
| GRU | 64 | Real |  |  |  |  |  |
| Fake |  |  |  |
| RNN | 128 | Real |  |  |  |  |  |
| Fake |  |  |  |
| LSTM | 128 | Real |  |  |  |  |  |
| Fake |  |  |  |
| GRU | 128 | Real |  |  |  |  |  |
| Fake |  |  |  |
| RNN | 256 | Real |  |  |  |  |  |
| Fake |  |  |  |
| LSTM | 256 | Real |  |  |  |  |  |
| Fake |  |  |  |
| GRU | 256 | Real |  |  |  |  |  |
| Fake |  |  |  |

3.(测试集上评价模型)根据表2中的结果,选择在验证集上效果最好的RNN, LSTM和GRU模型，并在测试集上评价这些模型。然后在测试集上评价得到accuracy、recall, F1，并计算得出Accuracy和AUC评价结果，填入表3。

Table 3. results of intermediate tasks on testing set

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Model | Hidden | Label | Precision | Recall | F1 | Accuracy | AUC |
| RNN |  | Real |  |  |  |  |  |
| Fake |  |  |  |
| LSTM |  | Real |  |  |  |  |  |
| Fake |  |  |  |
| GRU |  | Real |  |  |  |  |  |
| Fake |  |  |  |

1. 建议使用https://github.com/Embedding/Chinese-Word-Vectors中的微博语料库，也可以使用其他。 [↑](#footnote-ref-1)