**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3。

Table 3. statistics of data

|  |  |  |  |
| --- | --- | --- | --- |
|  | #real | #fake | #total |
| train.news.csv | 7844 | 2743 | 10587 |
| test.news.csv | 8659 | 1482 | 10141 |

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-0)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4。

Table 4. results of intermediate tasks on validation set

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Model | Hidden | Label | Precision | Recall | F1 | Accuracy | AUC |
| RNN | 64 | Real | 0.97 | 0.98 | 0.98 | 0.96 | 0.87 |
| Fake | 0.95 | 0.91 | 0.93 |
| LSTM | 64 | Real | 0.97 | 0.96 | 0.96 | 0.95 | 0.96 |
| Fake | 0.88 | 0.90 | 0.89 |
| GRU | 64 | Real | 0.97 | 0.99 | 0.98 | 0.96 | 0.98 |
| Fake | 0.96 | 0.90 | 0.93 |
| RNN | 128 | Real | 0.98 | 0.95 | 0.97 | 0.95 | 0.56 |
| Fake | 0.86 | 0.95 | 0.91 |
| LSTM | 128 | Real | 0.97 | 0.97 | 0.97 | 0.96 | 0.96 |
| Fake | 0.92 | 0.91 | 0.91 |
| GRU | 128 | Real | 0.99 | 0.94 | 0.96 | 0.95 | 0.67 |
| Fake | 0.85 | 0.96 | 0.90 |
| RNN | 256 | Real | 0.98 | 0.98 | 0.98 | **0.97** | 0.93 |
| Fake | 0.93 | 0.95 | 0.94 |
| LSTM | 256 | Real | 0.99 | 0.97 | 0.98 | **0.96** | 0.72 |
| Fake | 0.90 | 0.96 | 0.93 |
| GRU | 256 | Real | 0.99 | 0.97 | 0.98 | **0.97** | 0.72 |
| Fake | 0.91 | 0.96 | 0.93 |

3.(测试集上评价模型)根据表4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。在测试集上评价得到两种类别的accuracy、recall, F1，并计算得出Accuracy和AUC评价结果，填入表5。

Table 5. results of intermediate tasks on testing set

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | Label | Precision | Recall | F1 | Accuracy | AUC |
| RNN | Real | 0.95 | 0.96 | 0.95 | 0.92 | 0.84 |
| Fake | 0.75 | 0.71 | 0.73 |
| LSTM | Real | 0.94 | 0.96 | 0.95 | 0.91 | 0.93 |
| Fake | 0.73 | 0.66 | 0.69 |
| GRU | Real | 0.95 | 0.96 | 0.95 | 0.92 | 0.67 |
| Fake | 0.74 | 0.68 | 0.70 |

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-0)