### 一、本周研究内容

- **1.** 通过阅读文献,发现对 semi-supervised learning 中 self training (pseuo label) 应该是使用 labeled dataset 作为 validation set,对上周的实验进行了调整。
- 2. 整理了美国地区的 twitter 数据。

## 二、项目实施当前状态

目前正在实施 web app 后台数据的以及设定相关预警规则。

#### 三、本周成果

#### 1. semi-supervised 验证集的选择

阅读了相关文献, 在我发的 paper2, paper3 中,都使用了 labeled dataset 作为 validation set 而不是我上周使用的 pseuo labelled dataset。

# 4. Experiments

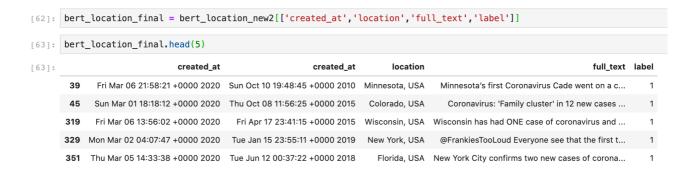
## 4.1. Handwritten Digit Recognition (MNIST)

MNIST is one of the most famous dataset in deep learning literature. For comparison, we used the same semi-supervised setting with (Weston et al., 2008; Rifai et al., 2011b). We reduced the size of the labeled training set to 100, 600, 1000 and 3000. The training set has the same number of samples on each label. For validation set, we picked up 1000 labeled exam-

图一: paper3 中相应截图

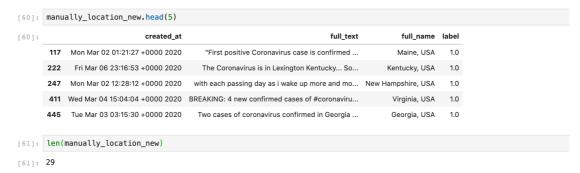
## 2. Web app 后台数据

使用 BERT 预测的结果以及手工标注的数据集,整理了地理位置是美国各州的全部数据。其中 BERT 预测的有756 个, 手工标注的有29 个。



图二: BERT 预测数据中美国各州的数据

# Yiming Zhang Weekly Report 7.10



图二: 手工标注数据中美国各州的数据

# 四、本周问题

美国各州的数据只有700多条,数据量比较少。

# 五、下周计划

完成 Web App 全部工作。