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。

```
[62]: bert_location_final = bert_location_new2[['created_at','location','full_text','label']]

[63]: bert_location_final.head(5)
```

[63]:

	created_at	created_at	location	full_text	label
39	Fri Mar 06 21:58:21 +0000 2020	Sun Oct 10 19:48:45 +0000 2010	Minnesota, USA	Minnesota's first Coronavirus Cade went on a c...	1
45	Sun Mar 01 18:18:12 +0000 2020	Thu Oct 08 11:56:25 +0000 2015	Colorado, USA	Coronavirus: 'Family cluster' in 12 new cases ...	1
319	Fri Mar 06 13:56:02 +0000 2020	Fri Apr 17 23:41:15 +0000 2015	Wisconsin, USA	Wisconsin has had ONE case of coronavirus and ...	1
329	Mon Mar 02 04:07:47 +0000 2020	Tue Jan 15 23:55:11 +0000 2019	New York, USA	@FrankiesTooLoud Everyone see that the first t...	1
351	Thu Mar 05 14:33:38 +0000 2020	Tue Jun 12 00:37:22 +0000 2018	Florida, USA	New York City confirms two new cases of corona...	1

图二: BERT 预测数据中美国各州的数据

```
[60]: manually_location_new.head(5)
```

[60]:

	created_at	full_text	full_name	label
117	Mon Mar 02 01:21:27 +0000 2020	"First positive Coronavirus case is confirmed ...	Maine, USA	1.0
222	Fri Mar 06 23:16:53 +0000 2020	The Coronavirus is in Lexington Kentucky... So...	Kentucky, USA	1.0
247	Mon Mar 02 12:28:12 +0000 2020	with each passing day as i wake up more and mo...	New Hampshire, USA	1.0
411	Wed Mar 04 15:04:04 +0000 2020	BREAKING: 4 new confirmed cases of #coronaviru...	Virginia, USA	1.0
445	Tue Mar 03 03:15:30 +0000 2020	Two cases of coronavirus confirmed in Georgia ...	Georgia, USA	1.0

```
[61]: len(manually_location_new)
```

[61]: 29

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。