项目计划

‘COVID-19 Twitter Mining and Text Classification for Epidemic Diseases Early-warning System’

—

2020.5.26

项目的背景

COVID-19全世界大爆发，尽管WHO已经进行了疫情相关的预警，但是这些预警还是不够及时。本项目拟从另一角度去建立一个疫情预警系统，即通过社交媒体的大数据与人工智能技术相结合。具体来说，本项目通过社交媒体twitter去搜集与COVID-19有关的数据，使用机器学习和深度学习去建立一个COVID-19预警分类器，将twitter分为 COVID-19在本地区潜在爆发相关的和无关系的这两类。然后，本项目提出一个新的预警系统的预警公式，针对全世界每个地区，根据分类到的twitter数量等，划分相应预警等级。并最终建立一个web app，将各个模块整合起来，建立一个完整的预警系统。

项目Pipeline

详情参考项目**思维导图**，项目一共分为8部分，分别为 twitter raw dataset, data preprocessing, twitter mining, annotation, text classification algorithms, evaluation, build early-warning system, paper writing。项目的重点在**text classification algorithms**和 **build early-warning system**。

Contribution

- **创建了一个标注好的与COVID-19相关的，可用于文本分类的twitter数据集。**
- **建立了一个新的text classification model，在本数据集上表现出色。**
- **提出了一个疫情预警的公式，并基于此建立了一个COVID-19疫情预警系统的web app。**

Deadline

论文拟投稿ELSEVIER的‘Technology Forecasting & Social Change’特刊上，本项目初稿完成时间due为8月初，然后交给美国的Zhuo Chen教授进行公共卫生方面的补充与修改，论文最终due是10月份。

任务划分

- **我的任务：**
  - 继续标注数据，完成对全部数据的标注。**（5.26 - 6.2）**
  - 解决数据imblanaced class的问题，以及使用文本增强对数据集进行扩充。**(6.2 - 6.9)**
  - Text classification model实现CNN、RNN，并在数据集上实验。**(6.9 - 6.16)**
  - **一起合作**完成新的model，model可能是基于CNN的改进或基于BERT的model，进行实验。**(6.16 - 7.7)**
  - 负责Web app的rule 设定，以及前端的相关工作。**(7.7 - 7.21)**
  - 写论文。**(7.21 - 8.4)**

- **Ke Chen同学的任务**：
    - 阅读我打包好的相关文献，一共15个。15个文档分别涵盖了项目申请书、创新点、预警系统架构的相关文献、twitter classification的相关文献、twitter数据集的相关文献、text classification的相关文献、我在周报中汇报的项目的目前进展以及项目整体思维导图。遇到不懂的地方可以及时问我，也希望Ke Chen同学自己也找一些相关文献，有更好的思路以及发现我哪里思路不对提出一些建议。**（预计 5.26 - 6.2）**
    - 标注数据集并且更新打标guidelines：整理后的数据集为文件 ‘coronavirus_merged_label.pkl’，一共包含7068个tweets。我目前已经打标了近一半数据，但是一个人打标的时候难免会太主观，所以希望Ke Chen同学在阅读完文献后，对项目整体理解了之后，也重新对数据集进行打标。**(预计 6.2 - 6.16)**
    - 一起合作完成新的model，model可能是基于CNN的改进或基于BERT的model，进行实验。**(预计 6.16 – 7.21)**
    - 负责Web app后端的相关工作。**(预计 7.21 – 8.4)**