Twitter 预警数据集

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

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。

任务描述

Twitter 预警系统旨在识别分类 twiiter 中的内容信息，输入一条 twitter 的文本内容，然后将 twitter 分为以下两个类别之一：与 COVID19 在本地潜在爆发相关、与 COVID19 在本地潜在爆发无关。数据集的两个类别举例如下表所示：

分类类别	Twitter 疫情数据集
与潜在爆发有关	**“I've been fighting a cold. Maybe I have the coronavirus. Lol”**
与潜在爆发无关	**"Tokyo 2020 could be postponed amid coronavirus outbreak, Olympic minister suggests" https://t.co/RJkIgnYfFB**

表一： 数据集分类举例

数据集说明

本数据集是使用 Twitter API 并使用关键词 ‘coronavirus’ 搜索得到的数据，时间选取为 2020 年 1 月 22 日到 2020 年 3 月 11 日间的数据。通过上述搜集规则，搜取到的初始数据为 2576357 条。由于本项目需要 twiiter 有具体的地理位置信息，从而才能估计各地区的潜在爆发的风险等级，因此数据集删除了没有位置信息的 twiiter，不是英文的 twitter 以及转推的 twitter。最终数据集有 7064 条有效 twitter 用于实验研究。

本项目的文本分类任务属于机器学习/深度学习中的监督学习，需要将数据进行人工标注，然后让模型通过算法去学习。我们将全部 7064 条数据进行了人工标注，保存在 ‘COVID19_twitter_dataset.csv’文件里。CSV 文件一共包含六列：index, created_at, id_str, place, full_text, label。其中 index 为 twitter 在原始数据集中的编号，created_at 为该 twitter 的产生时间，id_str 为 twitter 的原始 id，place 为该 twitter 的地理位置信息，full_text 为 twitter 的文本内容（研究对象），label 为我们对 full_text 手工标注的结果，其中 1.0 为与潜在爆发有关，0.0 为与潜在爆发无关。示例如下：

index	created_at	id_str	place	full_text	label
474	Tue Mar 03 19:58:00 +0000 2020	1234931050360406018	{'id': '7929cea6bd5b32bd', 'url': 'https://api.twitter.com/1.1/geo/id/7929cea6bd5b32bd.json', 'place_type': 'city', 'name': 'Mumbai', 'full_name': 'Mumbai, India', 'country_code': 'IN', 'country': 'India', 'contained_within': [], 'bounding_box': {'type': 'Polygon', 'coordinates': [[[72.74484, 18.845343], [73.003648, 18.845343], [73.003648, 19.502937], [72.74484, 19.502937]]]}, 'attributes': {}}	"Tokyo 2020 could be postponed amid coronavirus outbreak, Olympic minister suggests" https://t.co/RJkIgnYfFB	0.0

表二： 数据集说明

在标注文本时，我们根据项目研究内容以及目的，制定了 3 个打标的准则 (guidelines), 所有的打标都遵循以下 3 个打标准则。

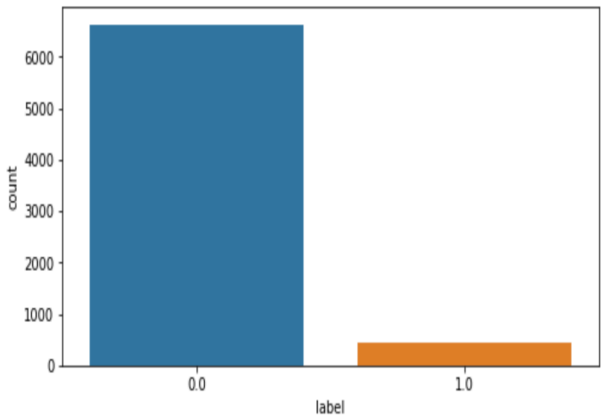
Annotation Guidelines	Examples in Dataset
1. Twitter 文本中有与 COVID19 相关症状的描述, 则被认为该数据为有关。 COVID-19 Symptoms including fever, dry cough, tiredness, aches, pains, nasal congestion, sore throat, diarrhea. Hence the first guideline is to label as positive if it describes such symptoms in the tweet. The descriptions can be formal (include keywords) or very colloquial.	I think I've caught a cold, unless it's Coronavirus... 🦠
2. Twitter 文本中涉及 COVID19 在本地区/国家确诊的。 If the tweet indicates the user or someone in his community/city/state has been tested positive for either viral test or antibody test, then this tweet should be considered positive.	"Not a hoax as #Trump said in SC rally. Illinois officials say patient has tested positive for #coronavirus https://t.co/lbgsYxqRo8"
3. Twitter 文本中涉及发现本地区有疑似 COVID19 的患者。 Someone may have COVID-19(suspect) or someone was infected or there is community spread, but not confirmed.	" A letter sent to the @PlymouthSch community warns that a student who just got back from Italy last month was hospitalized with flu-like symptoms. We're tracking this potential case of #coronavirus on @boston25 at 5 and 6:30 https://t.co/yZAswzS2dd"

表三： 文本标注规则及示例

数据集分布情况：7064 条 twitter 数据中，标注为 1 的(与潜在爆发相关)的有 434 条，标注为 0 的（与潜在爆发无关的）有 6630 条。

	label	full_text
0	0	6630
1	1	434

图一： 数据集的统计分布数据



图二·： 数据集分布柱状图

评价指标

本项目的评价指标为 Accuracy, Presicion, Recall, F1-score。

Accuracy= (TP+TN)/TP+FP+FN+TN

Precision=TP/TP+FP

Recall=TP/TP+FN

F1-score=2*Precision*Recall/Precision+ Recall