一、题目
灾难推文的自然语言处理

二、成员
2006500002 丁柳冉
2006500013 丁千惠
2006500025 刘霁莹

三、kaggle 链接
使用灾难推文进行自然语言处理 | 卡格尔 (kaggle.com)

四、科学意义

在当今社会，人们对信息的时效性、即时性和准确性的需求越来越高，尤其是在灾难事件方面。twitter 是一个拥有 1.86 亿日活跃用户的微博客服务网站，也成为了突发紧急情况时人们的重要沟通渠道。在灾难事件中，及时有效的信息是非常关键的。然而，由于社交媒体的普及和传播速度，灾难事件的相关信息常常受到谣言、虚假信息和噪音的干扰。因此，我们需要一种快速而准确地识别和分类推文的方法，以区分真实的灾难事件和虚假信息。

自然语言处理和机器学习技术可以帮助我们处理这些推文数据。为了帮助相关组织网络监测灾难发生及救援，我们可以通过建立一个机器学习模型预测 Twitter 推文发布灾难的真实性。这样的预测机器学习模型可以帮助将人们从众多繁杂冗余的信息中分离获取到真正的灾难事件信息，并提供更有效的救援支持，以使用监督学习算法对推文进行分类，从而预测其真实性。

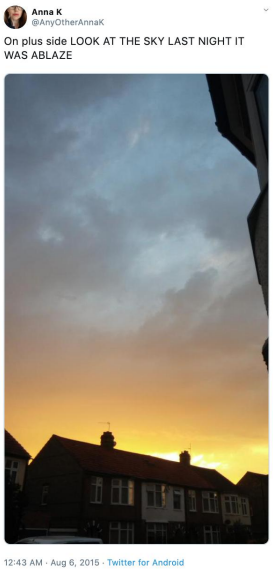
通过这种方式，我们可以更有效地监测灾难事件，提供及时准确的救援支持，并为相关组织提供有用的信息。这种方法不仅可以在灾难事件中发挥作用，还可以应用于其他领域的信息分类和监测任务中。

五、科学问题

随着互联网技术的不断发展和普及，人们的信息获取方式已经从传统的媒体渠道转变。在互联网时代，我们可以通过搜索引擎、社交媒体、在线新闻等方式获取到我们所需的信息，信息获取变得容易便捷。但同时，传统人工处理数据的方式已经无法适应当前大数据时代。如何基于自然语言处理技术实现机器自动将期望的数据判别出来是目前亟需研究的问题。

对于机器来说，它并不清楚一个用户发布的推文是否是真实的正在发生的灾难。例如：用户发送了“从正面看昨晚的天空，好像在燃烧一样”，作者明确使用了“ABLAZE”一词，但仅仅是为了形容火烧云，并非真正的火焰燃烧。这对于人类来说是显而易见的，但是对于机器来说便很难分辨该用户是否正在预告真实发生的火灾。

另外，考虑到信息快速传播的特点，许多错误、虚假和欺骗性信息也能够以迅雷不及掩耳之势散播开来，给人们的生命财产安全和公共安全带来严重威胁。因此，利用自然语言处理技术和机器学习方法进行推文分类和真假判别的研究具有重要的现实意义。在科学研究中，需要通过大规模数据集的构建、算法设计和性能评估等一系列工作逐渐逼近实际应用场景，完善推文分类系统的准确性，以帮助人们更好地了解灾难事件，及时采取行动，降低损失。



六、研究内容

1、数据集介绍

1.数据集文件

train.csv-训练集

test.csv-测试集

sample_submission.csv-正确格式的示例提交文件

2.数据集格式

训练集和测试集中都具有以下信息：①推特的文本（text）；②来自该推特的关键词（keyword）；③推特的发送地点（location）

3.数据集内容

①id：每条推特的唯一标识符

②text：推特的文本

③location：推特的发送地点（可能为空）

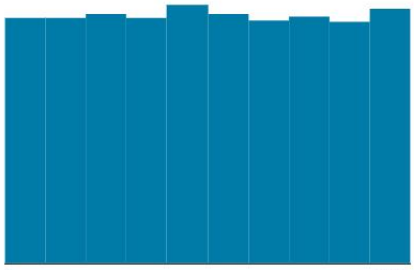
④keyword：推特中的特定关键词（可能为空）

⑤target：只在 train.csv 中存在，表示该条推特是否为真正的灾难（真为 1，假为 0）

4.数据集详情

①train.csv

id



Valid	7613	100%
Mismatched	0	0%
Missing	0	0%
Mean	5.44k	
Std. Deviation	3.14k	
Quantiles	1	Min
	2734	25%
	5408	50%
	8146	75%
	10.9k	Max

keyword

222
unique values

Valid	7552	99%
Mismatched	0	0%
Missing	61	1%
Unique	221	
Most Common	fatalities	1%

location

[null]	33%
USA	1%
Other (4976)	65%

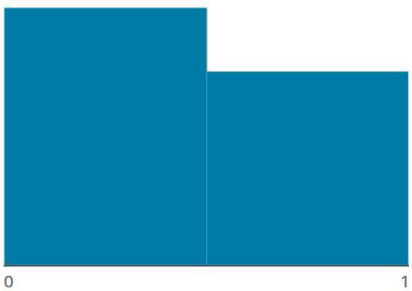
Valid	5080	67%
Mismatched	0	0%
Missing	2533	33%
Unique	3341	
Most Common	USA	1%

text

7503
unique values

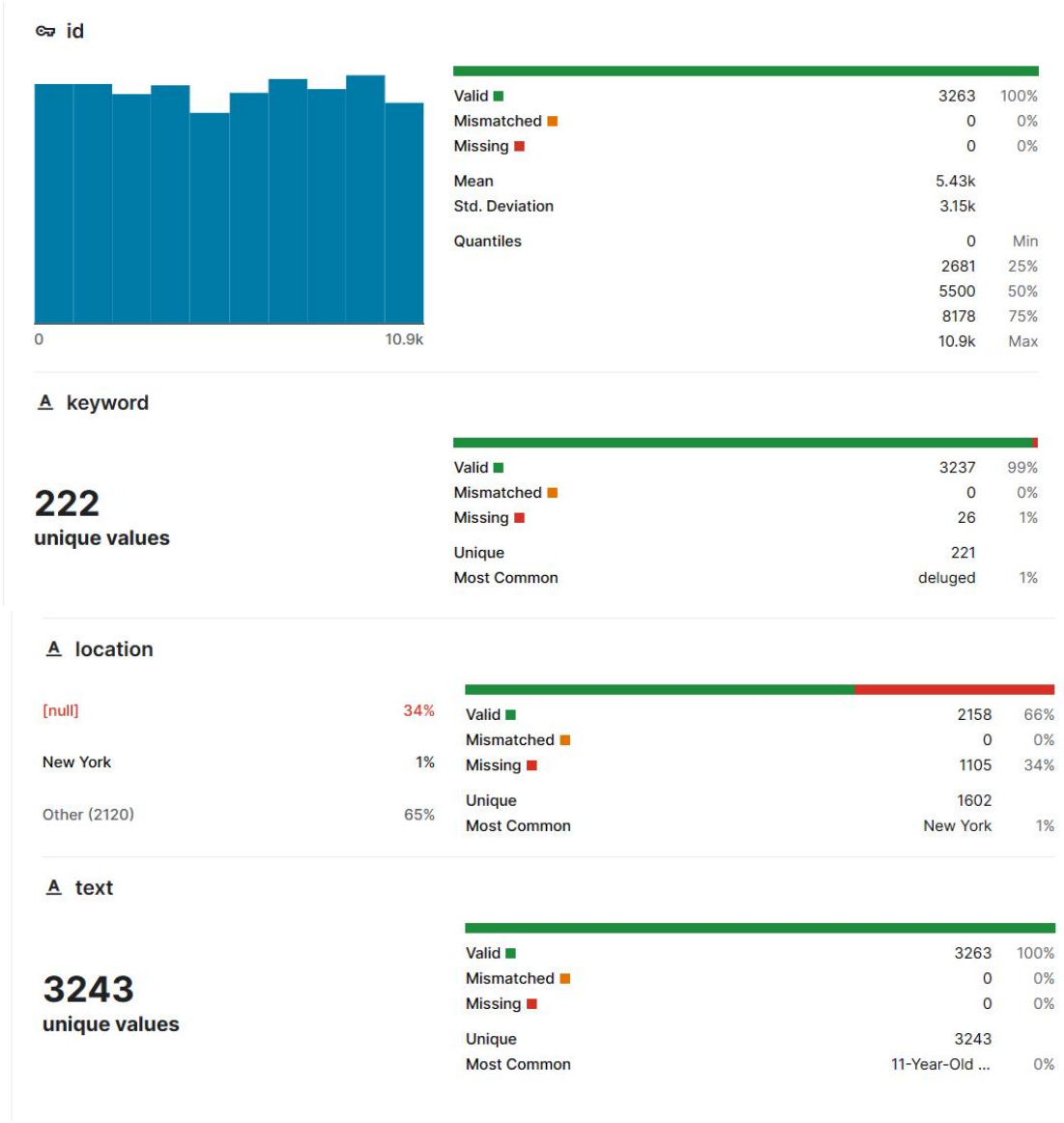
Valid	7613	100%
Mismatched	0	0%
Missing	0	0%
Unique	7503	
Most Common	11-Year-Old ...	0%

target



Valid	7613	100%
Mismatched	0	0%
Missing	0	0%
Mean	0.43	
Std. Deviation	0.5	
Quantiles	0	Min
	0	25%
	0	50%
	1	75%
	1	Max

②text.csv



2、技术路线

1.导入必要的包，包括 numpy、pandas、sklearn、nltk 等；

2.读取训练集和测试集的数据，存入 train 和 test 中；

3.定义清洗文本的函数 clean_text() 和去除非必须单词的函数 remove_stopwords()；

4.对训练集和测试集的文本数据进行清洗，得到干净的文本数据；

5.定义合并文本属性的函数 combine_attributes()，并使用 apply() 函数将其应用到 train 和 test 中；

6.使用 BertTokenizer 对干净的文本数据进行编码，得到训练集和测试集的编码数据；

7.使用 BertForSequenceClassification 建立分类器 clf，并用训练集的编码数据和标签训练分类器；

8.对测试集的编码数据进行分类预测，并计算模型的准确率。

七、预期目标

预测给定的推文是否是关于真正的灾难。如果是标签为 1；如果不是标签 0。

最终预测准确率达到 75%以上