**假新闻预测模型之建置及预测**

一、项目说明

假新闻（英语：Fake news）是通过传统新闻媒体（印刷和广播）或在线社交媒体，故意传播[错误](https://zh.wikipedia.org/wiki/%E9%8C%AF%E8%AA%A4%E8%B3%87%E8%A8%8A)或[恶作剧](https://zh.wikipedia.org/wiki/%E6%83%A1%E4%BD%9C%E5%8A%87)的信息，其目的是为了误导大众，以带来政治或经济利益，以及心理上的成就感。这些虚假信息通常是由记者因收受利益，进行报导时所引起的。因此，本项目要求考生开发先进的机器学习模型，以识别哪些文章可能会是假新闻。

二、假新闻的特征

假新闻一般有以下特征：

* [钓鱼式标题](https://zh.wikipedia.org/wiki/%E6%A0%87%E9%A2%98%E5%85%9A)
* [宣传](https://zh.wikipedia.org/wiki/%E5%AE%A3%E4%BC%A0)
* 讽刺性内容
* 垃圾内容
* 误导性的标题
* 有偏见的新闻

First Draft News则认为假新闻有以下特征：

* 讽刺性或模仿性内容
* 虚假连接
* 误导性内容
* 虚假上下文
* 冒名顶替的内容
* 操纵内容
* 捏造的内容

考生应用既有已经标定为真实新闻(Real News)及假新闻(Fake News)的数据，建置假新闻预测模型，以协助相关单位针对假新闻进行适当的风险处理，将假新闻所带来的影响降至最低。

三、数据

数据集将包含以下几个字段：

* id：新闻文章的唯一ID
* title：新闻文章的标题
* author：新闻文章的作者
* text：文章的文本；可能不完整
* label：将文章标记为可能不可靠的标签
  + 1：假新闻
  + 0：真新闻

四、评估方式

考生将数据利用10折交叉验证(10 Folds Cross-Validation)的方式来进行模型评估。模型的评估指针将以正确率(Accuracy)及假新闻的F1-Measure值来评估。

五、项目分析要求

考生须对所提供之数据进行以下的处理：

* 字根还原
* 词性标注 & N-Gram
* 非结构转结构-词袋模型
* 非结构转结构-词嵌入模型
* 以机器学习进行文本分类
  + Naive Bayes & Bayes Net
  + BP Neural Network
  + Support Vector Machine
  + Random Forest
  + Adboosting & xgboosting
* 以深度学习进行文本分类
  + CNN
  + RNN
  + LSTM & BLSTM

六、项目交付项目

* + 项目相关档案
  + 编程代码
  + 完整项目简报文件(ppt)

简报内容必需包含下列项目：

* + 项目目标达成情形说明
  + 项目进行步骤说明
    - 所有模型的处理过程
    - 参数调整与结果解释
    - 模型使用变量筛选步骤、方式及与目标变量间因果关系
    - 最终应用模型采用原因、方程式结果与入选变量解释
  + 分别描述模型排序后高低及相关解释

七、评分准则

|  |  |  |
| --- | --- | --- |
| 评分面向 | 项目 | 占比 |
| 1.工具使用 | 是否具有完整的文本挖掘处理流程；从数据处理到模型评估 | 20% |
| 各参数调整的合理性 |
| 是否有创意性的方法 |
| 2. 分析报告 | 是否充分了解项目目标 | 45% |
| 分析文本数据与目标轮廓分析 |
| 样本设计与数据分割处理方式 |
| 建模说明 |
| 数据准备、处理与转换能力 |
| 变量筛选、缩减方式与结果解释 |
| 预测模型使用方法的合理性与业务解释性 |
| 模型评估方式、变量重要性说明与结果解释 |
| 是否有进行进阶分析 |
| 3.进阶分析与创意 | 是否具有独创性的分析想法或产出 | 15% |
| 是否有提升排序力 |
| 4.排序力 | 正确率(Accuracy)及假新闻的F1-Measure值 | 20% |