**面向数据安全治理的数据内容智能发现与分级分类**

2020000885

胡帆

# 1. 问题介绍

## 1.1 问题名称

面向数据安全治理的数据内容智能发现与分级分类

## 1.2 题目背景：

随着企业信息化水平的不断提高，数据共享与开放对企业发展的作用日益凸显，数据已成为重要生产要素之一，企业在产业与服务、营销支持、业务运营、风险管控、信息披露和分析决策等经营管理活动中涉及到大量的业务数据，其中可能会包含企业的商业秘密、工作秘密，以及员工的隐私信息等，若因为使用不当，造成数据泄露，则有可能造成巨大的经济损失，或在社会、法律、信用、品牌上对企业造成严重的不良影响。同时，在合规要求层面，围绕数据安全，国家近年密集颁布《网络安全法》、《民法典》、《数据安全法》（征求意见稿）、《个人信息保护法》（征求意见稿）等，从国家法律层面强调对关键基础设施、各类APP应用中的敏感数据保护要求。而为了有效、规范保护企业敏感数据，其首要问题是对数据进行分级分类，以识别敏感数据，从而进一步围绕保护对象的全生命周期进行开放、动态的数据安全治理，解决数据开放共享与数据隐私保护的矛盾与统一。

现有的敏感数据识别与分级分类已广泛采用基于自然语言处理的语义识别技术，但会存在以下问题：

1. 需要有大批量、高质量的标注数据，花费大量的人力和时间，建设成本高。
2. 泛化能力不足，对新业务数据的适应能力弱，敏感数据的误报率和漏报率高。
3. 不能进行自我优化、自我学习，需要业务和技术领域专家共同进行人工干预，建设难度大。

## 1.3 问题任务：

识别样本中的敏感数据，构建基于敏感数据本体的分级分类模型，判断数据所属的类别以及级别。

1. 利用远程监督技术，基于小样本构建文档分类分级样本库。
2. 结合当下先进的深度学习和机器学习技术，利用已构建的样本库，提取文本语义特征，构建泛化能力强且能自我学习的文档分类分级模型。

### 1.3.1 数据与评测标准

1. 数据简介
2. 已标注数据：共7000篇文档，类别包含7类，分别为：财经、房产、家居、教育、科技、时尚、时政，每一类包含1000篇文档。

|  |  |  |
| --- | --- | --- |
| **字段信息** | **类型** | **描述** |
| id | string | 数据id |
| class\_label | string | 文本所属类别 |
| content | string | 文本内容 |

1. 未标注数据：共33000篇文档。

|  |  |  |
| --- | --- | --- |
| **字段信息** | **类型** | **描述** |
| id | string | 数据id |
| content | string | 文本内容 |

1. 分类分级测试数据：共20000篇文档，包含10个类别:财经、房产、家居、教育、科技、时尚、时政、游戏、娱乐、体育。

|  |  |  |
| --- | --- | --- |
| **字段信息** | **类型** | **描述** |
| id | string | 数据id |
| content | string | 文本内容 |

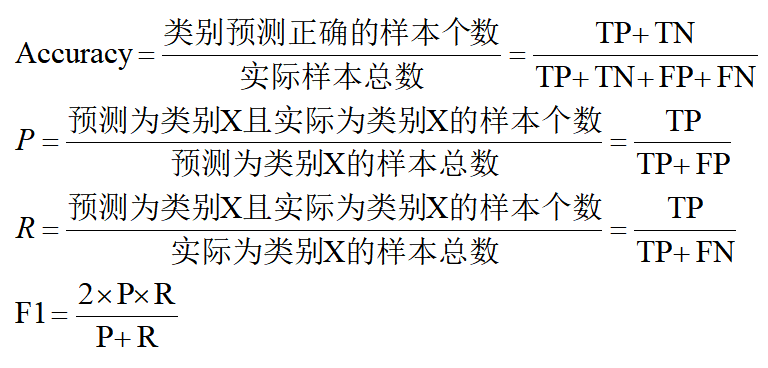
1. 分级信息

假设文档类别与文档级别有如下对应关系：

|  |  |
| --- | --- |
| **文档类别** | **文档级别** |
| 财经、时政 | 高风险 |
| 房产、科技 | 中风险 |
| 教育、时尚、游戏 | 低风险 |
| 家居、体育、娱乐 | 可公开 |

1. 评测标准

本赛题采用分类任务的精确率 P（precision）、召回率 R（recall） 和 F1 值三个指标作为模型性能的评判标准。



# 2. 算法方案

## 2.1 数据预处理方案

1. 核心问题分析："游戏、娱乐、体育" 这3个类别在测试数据中存在，但不在已标注的训练数据中，需要通过一些强规则从未标注数据中进行提取标注。考虑两种标注思路：
2. 数据中出现某类别关键词，即标注为该类别；
3. 仅当数据中只出现某关键词而未出现其他关键词时，才标注为该类别。
4. 针对上述两种标注方案分别在7个训练数据中已标注的类别上进行验证比较，选择最优方案。
5. 数据中出现某类别关键词，即标注为该类别

|  |  |
| --- | --- |
| Class\_label | 包含关键词且标注正确 |
| 家居 | 0.460 |
| 房产 | 0.116 |
| 教育 | 0.761 |
| 时尚 | 0.388 |
| 时政 | 0.003 |
| 科技 | 0.110 |
| 财经 | 0.051 |

1. 仅当数据中只出现某关键词而未出现其他关键词时，才标注为该类别

|  |  |
| --- | --- |
| Class\_label | 包含关键词且标注正确 |
| 家居 | 0.457 |
| 房产 | 0.595 |
| 教育 | 0.696 |
| 时尚 | 0.925 |
| 时政 | 0.333 |
| 科技 | 0.882 |
| 财经 | 0.745 |

从两表比较可知，标注方案二更加合理。

1. 从未标注数据中提取3种目标类标注数据时发现仅通过上述方案二进行标注获得的文档数目较少：

|  |  |
| --- | --- |
| Class\_label | 文档数目 |
| 体育 | 801 |
| 娱乐 | 641 |
| 游戏 | 2254 |

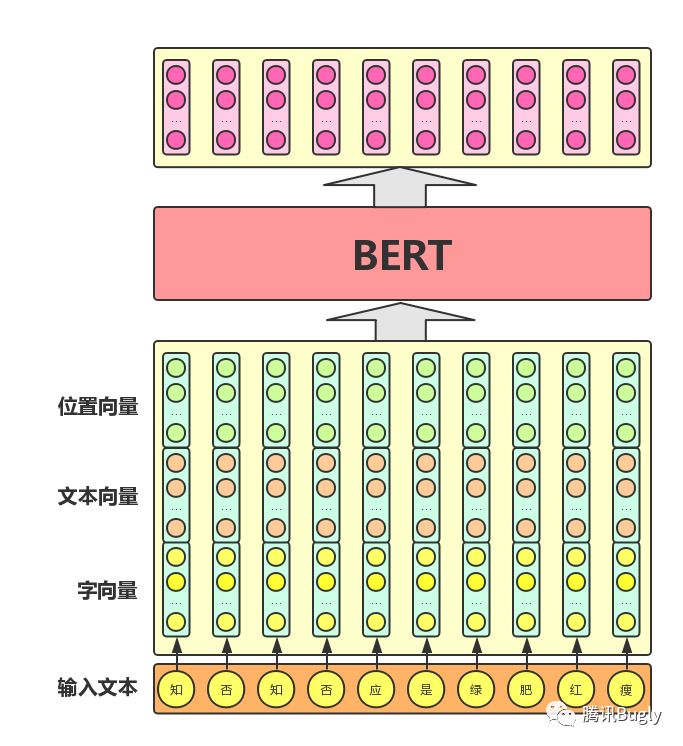
考虑通过TFIDF方法扩充关键词表：使用TFIDF方法筛选出上述3类目标类文档中前20个关键词，经过人工筛选后获得每类的关键词表。通过扩充的关键词表再次筛选目标类数据。

|  |  |  |
| --- | --- | --- |
| Class\_label | 关键词表 | 文档数目 |
| 娱乐 | '娱乐', '电影', '影片', '主演', '上映', '票房' | 1107 |
| 体育 | '体育', '篮板', '球员', '球队' | 2058 |
| 游戏 | '游戏', '玩家' | 3307 |

1. 将2中标注的3类数据与原训练数据中已标注的7类数据合并构成包含测试数据中10类数据的完整训练数据集。

## 2.2 模型结构

使用中文维基百科Bert 模型进行语义特征提取，然后使用加上全连接层进行分类：



训练方案中取每个训练文档前256字，使用中文维基百科预训练的BERT模型进行文本分类训练。

# 3. 提交成绩

如果未进行数据补充，仅仅采用已标注的7类进行训练，得到结果：



如果使用每篇文本的前256个字符作为输入，得到结果：



如果使用每篇文本的前512个字符作为输入，得到结果：



可以看到，前256个字符输入模型，结果最好，这也是最终比赛成绩：



# 4. 方案局限性反思

在从未标注数据中筛选标注3类不存在于已标注训练数据的目标类数据时，仅使用关键词表作为标注依据具有一定的局限性，实际上简单地筛选文档中出现的关键词不能完全代表文档应属类别。

例如，对于下面的文本信息：

-------------------------------------------------------------------------------------

这幅由摄影师AnnieLeibvitz所拍摄的照片中，这位美国游泳名将和前苏俄体操运动员LarisaLatynina坐于沙发上谈笑风生，并于8月16正式发布。 品牌称，这对组合有很多话语可谈，尤其是关于运动的征程。-他们来自不同的背景但是他们共享了通向荣誉的道路；他们各自最辉煌的时刻和记忆；他们各自的旅程等。品牌选择这对闻名世界的体育之星正因为他们很好解读了品牌的信仰-“传达激情于他人”。 广告注解说道：“2个不同的伟大旅程，以相同的道路到达顶点。LarisaLatynina和MichaelPhelps，体操与泳坛传奇。”继拳击名将MuhammadAli，演员AngelinaJolie，滚石乐队吉他手KeithRichards和U2领唱Bono后，这对组合将追随前者们的脚步登上这个核心旅游系列的画面。-------------------------------------------------------------------------------------

根据内容，这段文本应被分类为“娱乐”，但由于其中出现了大量与“体育”相关的关键词，如“运动，游泳，体操”等，在基于关键词表的标注方案中将会被分类为“体育”。这样造成标注错误，影响最终模型准确率。