新闻数据分析报告

摘要：

多类文本分类问题是自然语言处理中的重要问题，本报告通过对新闻文本数据的分析，建立词袋模型，TF-IDF 模型，和循环神经网络模型，进行新闻类别分类。

1 背景介绍

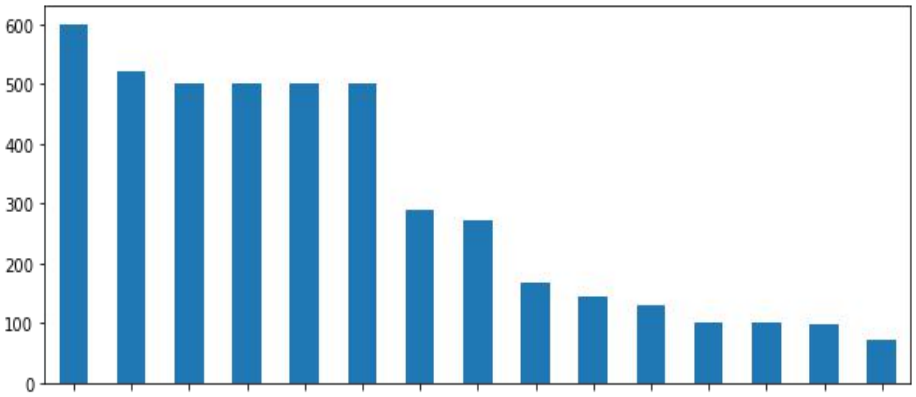
随着新媒体的发展，每天每分每秒都有大量新闻产生，新闻的来源也多种多样，比如 CCTV、今日头条等。当用户打开新闻应用后，呈现的是已经分类的新闻，应用还可以根据用户预先选择的偏好和历史浏览记录，为用户定制新闻的排列顺序，发送推荐内容到用户邮箱。这些功能的实现离不开数据分析的技术，本报告将根据提供的新闻数据，应用贝叶斯方法和 RNN 方法，来识别新闻的类别。

2 数据

2.1 新闻数据

我们一共有 4492 条新闻数据，其中每一条新闻分别有 7 个特征，分别是标题、标题链接、brief、keywords、发布时间、新闻类别、采集时的时间。所有数据都为文本数据。其中新闻类别包括 15 个类别，也是我们分类任务的标签。由图 1 可看出，标签分布不均匀，部分类别的新闻较多，本文将对不均衡数据的处理作为下一步研究的方向。采集时的时间可以大致划分为三个时间段，包括 2021-04-24 16:42:20 到 16:43:35，以及 2021-04-24 16:53:36 到 2021-04-24 18:54:35，还有 2021-05-27 18:19:06。

图 1: 标签分布直方图



2.2 缺失值处理

在 4492 条新闻中，有 75 条新闻确实标题，9 条新闻缺失概要，120 条新闻缺失关键词，所有新闻都有新闻类别。因为要新闻类别分类，我们重点关注的变量有概要和关键词，还有新闻类别，所以我们对

概要和关键词的缺失值进行处理。由于标题是概要的总结，所以我们用概要替代缺失的概要数据。对于关键词，我们使用 TF-IDF 在概要数据中提取关键词，替代缺失的关键词信息。

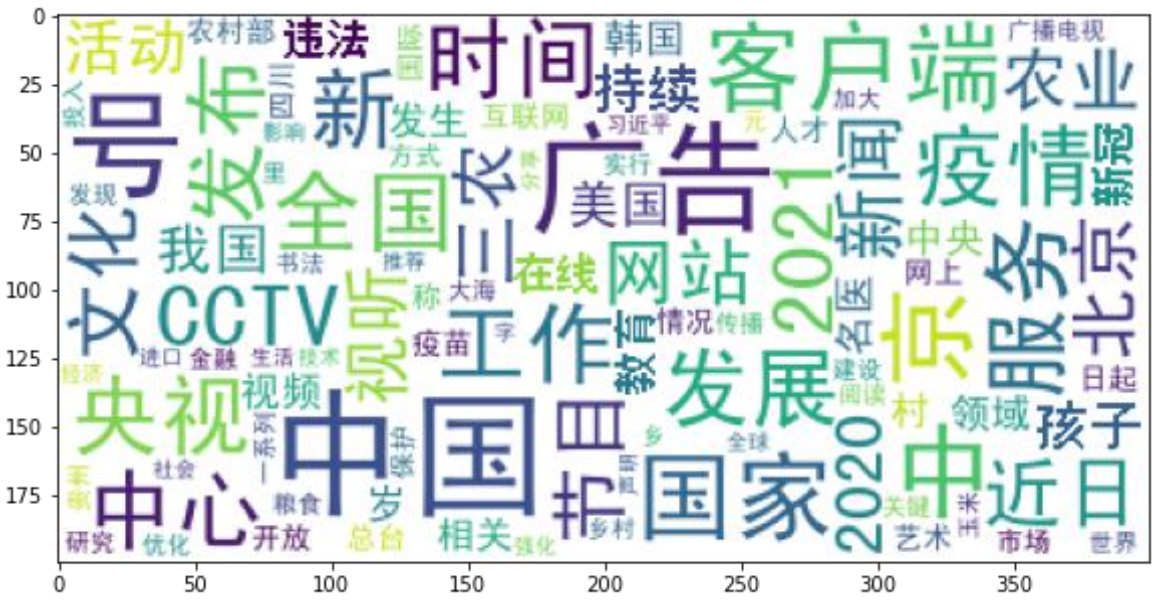
2.3 分词和数据清洗

首先对字符串进行分词处理，我们使用 jieba 分词，遍历存放新闻的列表，对每一篇新闻概要使用 jieba.lcut()进行分词，去掉字符个数小于等于 1 的词语和换行符，创建 dataframe 并存储。使用停用词字典数据进行数据清洗，去除不重要的词汇和标点符号。我们将舍弃出现在停用词字典中的词汇。停用词字典来自日常词汇收集，本报告停用词字典来自开源数据，但不同的文本数据常常需要不同的停用词字典，训练更适合新闻数据的停用词字典将是我们潜在的以后的研究方向。去掉停用词字典中词汇后，我们得到清理好的数据 df_brief，以及完整的词汇表 df_all_words。

2.4 词云图

我们使用全部词汇的数据绘制词云图，为了最好的展示效果，选择 100 个呈现目标，设置背景为白色，绘制的词云图如图 2。根据常识，很多词汇常常出现在日常的新闻中，比如“中国”等地区词汇，“2021”等时间词汇，“发展”等贴近民生的词汇等。

图 2：全部词汇词云图



2.5 提取关键词

TF-IDF 可以帮我们找到归宿文本数据里重要程度最高，最有价值的词。它的主要理念是，如果一个词在整个语料库中出现的次数都很高，那么这个词的重要程度就不高，或者可以成为通用词。如果另一个词在整体的语料库中的词频很低，但是在这一篇文章中却大量出现，就有理由认为它在这篇文章中很重要。我们这里以最后一条新闻为例，概要内容为“上周，两档团体偶像节目扎堆更新，《创造营 2021》和《青春有你 3》还首次在同一档期正面对抗，一次性上线 200 多位练习生的激烈场面，也再

次将国内偶像团体综艺节目竞争的白热化摆到了台前。”TF-IDF 提取的五个关键词为“偶像，团体，练习生，两档，2021”，新闻中关键词数据为“男团,偶像节目,代际危机”。虽然关键词并不完全一样，但关键词的含义大致一样，比如练习生团体和男团含义相似，偶像和偶像节目很类似，具体可以通过图神经网络等技术分析。已经得到了关键词，我们用它们来替代原始数据中缺失的关键词数据。

3 数据分析

3.1 词袋模型

为了方便计算机的处理，我们需要将文本数据转化为数字，用数字来表示词汇。首先对新闻类别做处理，用 1-15 的数字来匹配新闻类别，匹配对应关系如图 3 所示。

图 3：标签的映射

	国际	文娱	军事	三农	教育	书画	经济	健康	科技	人物	生活	社会	国内	农经	法治
mapping	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

由于每篇新闻概要字数不同，而机器学习要求输入数据的维度是统一的，我们建立语料库，根据每条新闻中词汇与语料库中词汇对应关系，将概要数据转换为同一维度。我们建立词袋模型，统计词频，得到文本特征。

首先划分训练集和测试集，这里我们使用 sklearn 中 train_test_split 进行划分，其中训练集包含 3369 条新闻，测试集包含 1123 条新闻。将数据转化成列表模式。3369 篇文章共创建的语料库的容量为 2 万多，而限制保留指定的特征后，我们选取词汇频数最大的 4000 个进行保留，以避免稀疏矩阵。然后词袋模型的特征建模，导入贝叶斯模型。接下来对测试集进行模型评测，我们得到 0.683。应用词袋模型对“2021 点 广西 防城港市 上思县 思阳 镇江 平村 路口 一辆 装满 沙子 货车 侧翻 司机 被困 请求 救援 接到 报警 上思县 消防 救援 大队 出动 车 名 消防 指战员 赶赴现场 施救”做预测，得到预测结果 12，对比图 3 得标签为“社会”，检查对应 keywords，预测正确。

3.2 TF-IDF 制作特征值

TF-IDF 模型能够给重要程度更高的词汇赋予更高的权重，应用 TF-IDF 模型进行特征建模，得分为 0.622。与使用词袋模型相比，得分下降。

3.3 关键词

上面我们用概论作为预测变量，来进行新闻类型分类。接下来我们使用关键词作为预测变量，应用相同的步骤，进行新闻类型分类。模型的结果是，使用词袋模型得分 0.478，使用 TF-IDF 模型得分 0.459。同样，词袋模型的表现优于 TF-IDF 模型，但总体模型表现不佳，其原因可能是关键词数据太少，很多新闻的关键词数量为 3 到 5，从而很难做出准确分类。

比如应用 TF-IDF 模型对“广西 上思 消防”做分类，我们模型的预测结果为“书画”，而实际标签为“社会”，模型预测错误。

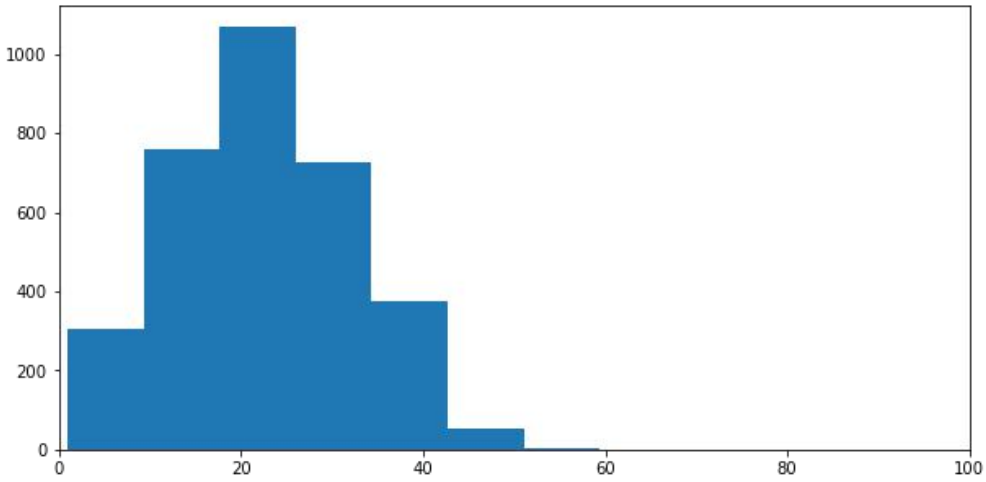
3.4 结合概要和关键词

由 3.3 的模型结果，我们考虑是否新闻包含词汇越多，模型表现越好。所以我们建立一个新的变量 words，结合概要和关键词中词汇。比如对第一条新闻，我们的新变量为“赣州 法院 审理 案件 中 诈骗 分子 假冒 海外 女 留学生 各类 社交 软件 境外 273 名 男性 实施 诈骗境外男性 被骗 电信 诈骗”。模型的结果是，使用词袋模型得分 0.690，使用 TF-IDF 模型得分 0.627。与使用词袋模型相比，TF-IDF 模型表现不佳。与使用概要变量比，使用新变量两个模型得分都有所增加，词袋模型增加幅度更大。

4 思考与讨论

为了进一步提高模型的得分，我们考虑更复杂的模型，比如循环神经网络。我们使用 TensorFlow 作为我们的机器学习框架。首先进行分词，我们使用 tensorflow.keras 中的 Tokenizer 进行分词，检查每一条新闻的数据长度，如图 4 所示。词汇的长度大多在 0 到 50 词之间，有一些新闻词汇很多，但这里我们把长新闻截断，选择最长词汇量为 50。

图 4: 词汇长度



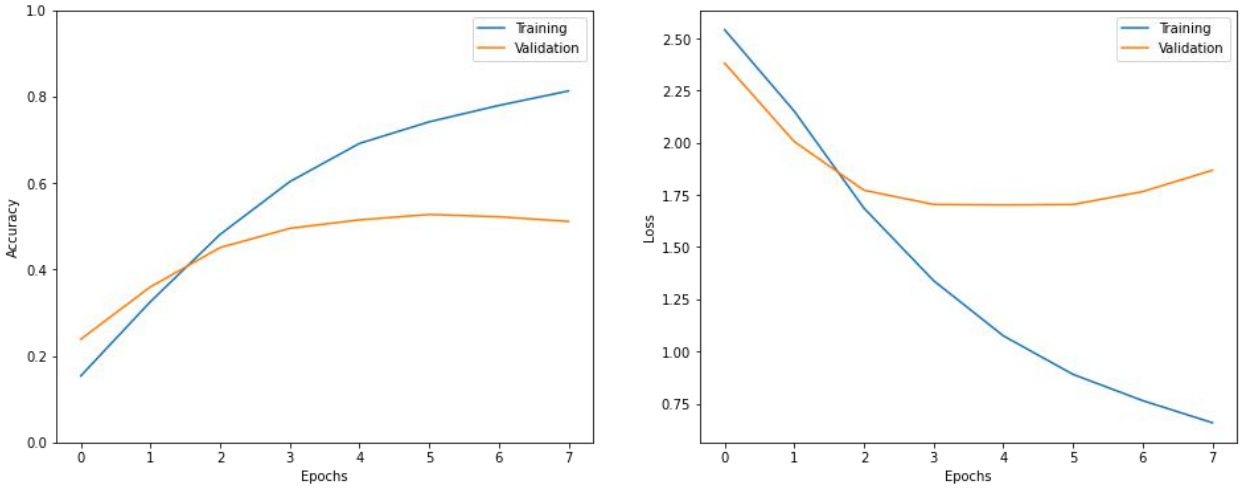
接着建立模型，RNN 模型参数如图 5.

图 5: RNN 模型总结

```
Model: "sequential_7"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_14 (Embedding)     (None, 50, 16)            64000
_________________________________________________________________
bidirectional_14 (Bidirectio (None, 50, 40)            5920
_________________________________________________________________
bidirectional_15 (Bidirectio (None, 40)                9760
_________________________________________________________________
dense_7 (Dense)              (None, 16)                656
=================================================================
Total params: 80,336
Trainable params: 80,336
Non-trainable params: 0
_________________________________________________________________
```

我们使用测试集作为 validation set，训练模型。模型经过 8 次迭代后趋于稳定。我们使用 accuracy 和 loss 来模型评估，结果如图 6.尽管模型在训练集上获得了很高的准确率，但在 validation set 上，模型表现还是不尽人意。

图 6：模型评测



5 结论

本报告分析了新闻文本数据集，通过分词、数据清洗、特征清洗等步骤，建立词袋模型、TF-IDF 模型和 RNN 模型，进行文本分类。我们发现，通过结合概要和关键词，应用词袋模型，我们能得到最佳的模型评分 0.690。但这个评分不理想，依然有很大的提升空间。模型的表现很可能与标签的分布不平

衡有关。同时我们还可以将标题等数据加入模型，进行文本分类。同时我们发现标题链接中存在特殊文本，如 news、arts 等，未来我们可以应用这些数据进行新闻类别分类。

6 参考文献

https://www.coursera.org/projects/tweet-emotion-tensorflow

https://blog.csdn.net/lys_828/article/details/108990366