新闻分类任务报告

程智镝

2023 年 11 月 24 日

作业任务

每条数据包含属性: `category`, `headline`, `authors`, `link`, `short_description`, `date`，一共有 42 个分类。

a. 数据集描述

b. 数据处理流程

c. 分类模型选择，设计，训练，验证，测试

d. 限 python 或 java 实现模型

1 数据集描述

使用的数据集包含了新闻文章的各种属性，共有 42 个不同的分类。数据集中的内容包括 headline（标题）、authors（作者）、link（链接）、short_description（简介）、date（日期）等。经过数据处理，得到了训练集和测试集，并进行了统计分析。

2 数据处理流程

在数据处理阶段，对数据进行了清洗，包括处理缺失值等。然后将数据集分割为训练集和测试集，以便进行模型训练和测试。

3 分类模型选择及设计

在分类模型选择方面，选择了逻辑回归模型。使用了 TF-IDF 进行文本特征提取，并将提取出的特征作为逻辑回归模型的输入。使用 Python 中的 Scikit-learn 库来实现模型的训练、验证和测试。

4 模型训练和评估

在模型训练阶段，使用训练集对逻辑回归模型进行了训练。之后，对模型进行了评估，包括打印了混淆矩阵和分类报告。混淆矩阵展示了模型的预测结果和真实标签的对比，而分类报告包括了精确度、召回率和 F1 得分等指标，以及每个类别的支持数。

4.1 性能指标

模型在测试集上的性能指标如下：

Metric	Precision	Recall	F1-Score	Support
Accuracy	0.55	-	-	41782
Macro Avg	0.51	0.37	0.41	41782
Weighted Avg	0.54	0.55	0.52	41782

表 1: 模型性能评估

tips:”Accuracy” 表示准确率,”Macro Avg” 表示宏平均值,”Weighted Avg” 表示加权平均值。

混淆矩阵如下所示：

$$\begin{bmatrix} 43 & 15 & 7 & \dots & 0 & 1 \\ 15 & 27 & 6 & \dots & 14 & 1 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 1 & 3 & \dots & 5 & 198 \\ 0 & 0 & 4 & \dots & 6 & 37 \end{bmatrix}$$

svm，decisionTree: 另外还实现了 svm，决策树的模型分析，结果可运行查看

Category	Precision	Recall	F1-Score	Support
ARTS	0.34	0.13	0.19	324
ARTS & CULTURE	0.34	0.09	0.15	288
BLACK VOICES	0.45	0.31	0.36	867
BUSINESS	0.47	0.42	0.44	1175
COLLEGE	0.54	0.27	0.36	234
COMEDY	0.56	0.39	0.46	1085
CRIME	0.50	0.52	0.51	695
CULTURE & ARTS	0.74	0.24	0.37	218
DIVORCE	0.78	0.57	0.66	710
EDUCATION	0.41	0.21	0.28	191
ENTERTAINMENT	0.50	0.72	0.59	3417
ENVIRONMENT	0.53	0.16	0.25	270
FIFTY	0.49	0.12	0.19	296
FOOD & DRINK	0.59	0.62	0.61	1313
GOOD NEWS	0.40	0.13	0.20	305
GREEN	0.35	0.26	0.30	504
HEALTHY LIVING	0.31	0.19	0.24	1295
HOME & LIVING	0.69	0.64	0.66	891
IMPACT	0.38	0.20	0.27	674
LATINO VOICES	0.58	0.18	0.28	209
MEDIA	0.55	0.34	0.42	568
MONEY	0.49	0.26	0.34	368
PARENTING	0.49	0.56	0.52	1785
PARENTS	0.40	0.23	0.29	715
POLITICS	0.63	0.84	0.72	7045
QUEER VOICES	0.73	0.60	0.66	1259
RELIGION	0.57	0.40	0.47	502
SCIENCE	0.61	0.34	0.44	459
SPORTS	0.63	0.56	0.59	1018
STYLE	0.47	0.16	0.23	462
STYLE & BEAUTY	0.66	0.74	0.70	1894
TASTE	0.38	0.12	0.18	421
TECH	0.58	0.35	0.44	458
THE WORLDPOST	0.48	0.36	0.41	765
TRAVEL	0.60	0.71	0.65	1970
U.S. NEWS	0.31	0.05	0.09	264
WEDDINGS	0.78	0.69	0.74	766
WEIRD NEWS	0.37	0.21	0.27	562
WELLNESS	0.45	0.71	0.55	3677
WOMEN	0.41	0.30	0.35	694