大数据智能分析理论与方法

数管 2302 林子晗 2023113223

2025 年 1 月 7 日

目录

1 摘要

摘要

本次作业针对文本分类任务中的数据预处理问题，提出了一套基于停用词清理和格式统一的处理流程，以提高模型对 AI 生成文本和人工文本区分的精度。通过对数据集的缺失值、重复值及无意义字符进行排查与清理，确保数据的完整性和一致性；利用自然语言处理工具库 NLTK 的英文停用词表以及自定义停用词列表，对文本中的停用词进行剔除，有效减少了数据冗余。生成标准化文本数据列，为后续特征提取及机器学习建模奠定了坚实基础。

2 介绍 Introduction

生成式 AI 的快速发展极大地推动了自然语言处理技术的进步，生成文本的质量已经能够高度接近人工创作。然而，这种技术的普及也带来了潜在的滥用问题，例如虚假信息传播、恶意自动生成评论等。因此，如何辨别 AI 生成的文本，成为一项具有理论与实践意义的重要课题。

本文提出了一套基于 TF-IDF 特征提取与随机森林分类的方案，通过数据清洗、去除停用词以及文本格式规范化，显著提升了分类效果。从数据质量和模型性能双重角度优化分类精度。结果不仅可服务于文本分类领域，还可为其他基于生成式 AI 的识别任务提供技术参考。

3 相关工作

3.1 早期识别方式

在生成文本的辨别中，早期研究主要依赖于基于规则的简单方法，如关键词匹配。然而，这类方法对于复杂句式和同义词替换的情况表现较差. 近年，机器学习模型与大语言模型 [2] 逐渐被广泛应用。这些方法利用语义特征和上下文信息，显著提高了检测性能。

3.2 数据预处理与特征提取

数据质量对文本分类结果影响重大, 这次作业我使用以下方法进行数据处理

- 数据清洗：本次作业采用 pandas 库对无关字符（如空格，换行符）进行删除，对重复文本进行删除，统一文本格式。
- 停用词剔除：使用标准停用词库（NLTK）对文本中的停用词进行剔除，消除无意义词对结果的干扰。
- 高频词提取：使用 TF-IDF 方法对核心特征进行提取，转化特征矩阵。

3.3 机器学习模型的使用

随机森林作为一种强大的集成学习方法，因其良好的泛化能力和高效处理高维数据的优势，广泛应用于文本分类任务 [1]. 通过构建多个决策树并结合其投票结果，随机森林在处理复杂模式时有显著优势，尤其适用于具有非线性关系和高维特征的文本数据。

本次在特征工程阶段采用 TF-IDF 方法，有效捕捉文本的词频信息，并利用该方法将高维稀疏空间转化为较为密集且具有代表性的特征向量。采用分层 k-fold 交叉验证对随机森林模型进行评估，有效避免了由于数据不均衡引起的训练偏差，提升了模型的泛化能力。

4 模型方法

4.1 基本原理

随机森林的最终分类输出是所有树的投票结果，决策规则可以表示为：

$$\hat{y} = \text{Majority Voting}(T_1(x), T_2(x), \ldots, T_n(x))$$

其中，$\hat{y}$ 是最终的分类标签，$T_i(x)$ 表示第 i 棵树在输入数据 x 上的预测结果，n 是树的总数。

4.2 特征选择与预处理

在构建随机森林模型前，需要对原始数据进行特征选择和预处理。本次作业采用 TF-IDF 算法进行高频词提取，其中，TF-IDF 权重的计算公式如下：

$$\text{TF-IDF}(t, d) = \text{TF}(t, d) \times \text{IDF}(t)$$

$$\text{TF}(t, d) = \frac{\text{词频}(t, d)}{\sum_{t'} \text{词频}(t', d)}$$

$$\text{IDF}(t) = \log \frac{N}{\text{文档包含}t\text{的总数}}$$

其中，$\text{TF}(t, d)$ 表示词 t 在文档 d 中的词频，$\text{IDF}(t)$ 表示词 t 的逆文档频率，N 是文档总数。TF-IDF 的乘积可以有效衡量某个词对于给定文档的重要性，并帮助随机森林模型更好地理解文档内容。

5 总结

本次作业通过对 AI 生成文本的检测问题进行研究，获得了以下几点主要结果：

- 特征提取与数据预处理的方法：TF-IDF 和停用词的剔除增强了模型对重要信息的敏感性，避免了模型受到无意义词汇的干扰。
- 随机森林模型的应用：通过特征提取方法所训练的随机森林模型对于相关问题有较强可靠性，预测结果准确率较高。

本研究的不足在于模型的可解释性较低，随机森林作为一个黑箱模型，很难直接追溯每一个预测决策的具体原因。在未来的研究中，可以尝试结合更多的模型方法，进一步提升 AI 生成文本分类的准确性。

未来的优化工作方向包括：

- 模型的改进与扩展：探索与深度学习结合的方法，特别是在预训练语言模型方面，例如利用 BERT 等神经网络模型提升文本理解和分类能力。
- 更多数据集的使用：在多个不同类型的数据集上进行实验，以进一步验证随机森林模型的适用性和稳定性。
- 可解释性增强：通过使用更为可解释的模型（如 XGBoost）或者附加可解释性工具来增加模型结果的透明度，使得分类决策更加容易被理解。

6 自己的总结

其实根本没干啥就随便弄了个模型。。也没做什么工作
这方面我还刚入门，需要继续学

参考文献

[1] 樊迪. 基于随机森林的翻译文本误译语句自动识别方法. 自动化技术与应用, 41(05):121–124, 2022.

[2] 熊思棋. 基于上下文语义的幽默文本识别和生成方法研究. 2023.