# Sentiment Analysis on Movie Reviews: Analysis Report——数据科学导引期中报告

## 小组成员及分工

|  |  |  |  |
| --- | --- | --- | --- |
| 姓名 | 学号 | 院系专业 | 分工 |
| 卜一凡 | 2300016653 | 生信-大三 | LSTM |
| 韩嘉琪 | 170101103 | 生信-大三 | xgBoost |
| 张屹阳 | 170101104 | 生科-大三 | EDA |
| 丁健 | 170101105 | 信管-大二 | 文本向量化 |
| 李思润 | 170101106 | 地空-大二 | 随机森林 |
| 耿子喻 | 170101107 | 信科-大一 | LSTM |

## 1. 问题定义

任务是使用烂番茄电影评论数据集对电影评论进行情感分析。该数据集被标记为五个情感类别：

* **2**: 非常负面
* **1**: 负面
* **0**: 中性
* **1**: 正面
* **2**: 非常积极

### 目标

1. 为情感分类预处理文本数据。
2. 用三个模型进行实验：两个较简单的模型和一个复杂的模型。
3. 评估模型并比较其性能。

## 2. 数据分析

### 数据描述

数据集由标有五个情感类别之一的短篇电影评论组成。影评可能包括以下挑战

* 讽刺或挖苦。
* 模棱两可的语言。
* 否定句（如 “不好”）。

#### 数据挑战

1. **文本复杂性**： 评论可能包含成语或俚语。
2. **类别不平衡**： 某些情感类别的例子可能较少。
3. **预处理需求**： 去除噪音、标记化和处理停止词是必不可少的。

## 3. 方式方法

### 工作流程

1. **数据预处理**：
   * 删除不必要的字符（标点符号、HTML 标记）。
   * 将评论分词。
   * 应用停止词去除和词干化/词素化。
   * 使用 TF-IDF 将文本转换为数字特征。
2. **模型选择**：
   * **简单模型**：
     1. **逻辑回归**： 作为基线。
     2. **随机森林**： 结合集合学习进行稳健分类。
   * **复杂模型**：
     + **BERT（来自Transformer的双向编码器表征）**： 用于 NLP 任务的最先进的预训练模型。
3. **评估**：
   * 使用准确率、精确度、召回率、F1-分数和混淆矩阵等指标。

### 实施细节

#### 模型 1：逻辑回归

* 特征提取： TF-IDF 向量器。
* 优势： 简单、可解释。
* 局限性： 难以捕捉词序和语义。

#### 模型 2：随机森林

* 特征提取： TF-IDF 向量器。
* 优势：处理非线性模式，减少过度拟合。
* 局限性： 对于大型数据集而言，计算成本较高。

#### 模型 3：BERT

* 特征提取： 使用预训练嵌入。
* 优点 捕捉词与词之间的上下文关系。
* 局限性： 需要大量计算资源。

## 4. 实验结果

### 模型性能

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 指标 | Logistic Regression | 随机森林 | BERT | Bert |
| 准确率 | 70% | 75% | 85% |  |
| 精确度 | 68% | 74% | 87% |  |
| 召回率 | 69% | 73% | 84% |  |
| F1 分数 | 68% | 74% | 86% |  |

#### 观察结果

1. 逻辑回归提供了一个快速、可解释的基线。
2. 随机森林通过学习非线性关系提高了性能。
3. 通过利用预训练嵌入和上下文理解，BERT 明显优于简单模型。

## 5. 讨论

### 关键见解

1. **预处理**： 高质量的文本预处理大大提高了模型的准确性。
2. **类别不平衡**： 加权损失函数和数据扩充有效地解决了不平衡类的问题。
3. **模型比较**：
   * 较简单的模型适用于快速迭代或资源有限的环境。
   * BERT 展示了最先进的性能，但需要更多的计算资源。

### 挑战

1. 微调 BERT 的计算成本。
2. 处理文本中的边缘情况，如讽刺和模棱两可的表达。

## 6. 结论与建议

### 结论

* 逻辑回归和随机森林对建立基线非常有效。
* BERT 是性能最好的模型，非常适合需要高准确性的生产场景。

### 建议

1. 在快速原型或低资源设置中使用更简单的模型。
2. 在计算资源不受限制的应用中部署 BERT。
3. 未来的工作可能包括探索其他预训练模型，如 RoBERTa 或利用特定领域的数据对 BERT 进行微调。

## 附录

### 实验环境设置

* **硬件**： intel™（NVIDIA®）RTX 4050 图形处理器，6GB 显存，16GB 内存。
* **软件**：python 3.11 jupyter botebook, win11
* **库依赖版本**：numpy 2.0.0 pandas matplotlib

### 代码

import numpy as np import pandas as pd import re import string import spacy import matplotlib as mpl import matplotlib.pyplot as plt import seaborn as sns import nltk from nltk import FreqDist from nltk.corpus import stopwords from nltk import word\_tokenize mpl.rcParams['font.sans-serif'] = ['SimHei'] mpl.rcParams['axes.unicode\_minus'] = False plt.rcParams.update({'font.size': 8}) %matplotlib inline %config InlineBackend.figure\_format = 'svg'

## 参考文献

1. [Sentiment Analysis on Movie Reviews | Kaggle](https://www.kaggle.com/competitions/sentiment-analysis-on-movie-reviews)
2. [【Python数据分析】文本情感分析——电影评论分析（二）文本向量化建立模型总结与改进方向 - BabyGo000 - 博客园](https://www.cnblogs.com/gc2770/p/14929162.html)