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

一、引言

1. 选题分析

本次作业共提供了五个选题，GPT对其分析结果如下：

> **1. Sentiment Analysis on Movie Reviews**
>
> - **难度**： 中等到较高
> - **原因**： 该任务需要处理大量的文本数据，并进行情感分类。处理文本的难点在于要解决语言的多义性和复杂的语境问题，还需要对模型进行优化才能达到较高准确度。使用预训练模型如 BERT 可以提高效果，但对资源和模型调优的要求较高。
> - **推荐人群**： 有一定自然语言处理基础，或希望提升文本数据处理技能的小组。
>
> **2. Predict Health Outcomes of Horses**
>
> - **难度**： 中等
> - **原因**： 这是一个三分类任务，涉及健康数据的处理。分析健康数据的特点在于需要对特征工程有较好的理解，尤其是如何处理分类变量和连续变量。可以尝试使用树模型或神经网络，但不需要复杂的文本或图像处理。
> - **推荐人群**： 希望加强分类问题处理能力，尤其是面向结构化数据的小组。
>
> **3. Digit Recognizer**
>
> - **难度**： 低到中等
> - **原因**： 手写数字识别是机器学习中的经典任务，使用卷积神经网络（CNN）即可获得较好效果，且资源需求适中。数据集较为简单，不需要复杂的数据预处理，是入门级的计算机视觉任务。
> - **推荐人群**： 适合刚接触计算机视觉的同学，或者希望体验深度学习基础应用的小组。

> **4. ML Olympiad * Sustainable Urban Living**
>
> - **难度**： 较高
> - **原因**： 该任务涉及多种连续和离散特征的房产数据，且输出为连续变量，因此适合用回归模型进行预测。为提升 RMSE 评分，可能需要较深入的特征工程和模型调优。此外，房产数据可能具有较多的噪声和离群值，需花较多精力在数据预处理上。
> - **推荐人群**： 对回归问题或希望挑战数据预处理和特征工程的小组。
>
> **5. Tabular Playground Series * Clustering Project**
>
> - **难度**： 中等到较高
> - **原因**： 这是一个无监督的聚类任务，难度在于缺少标签和真实的类别数量，需要自行探索数据结构，确定合适的簇数和算法。选择合适的评价方法（如调整兰德指数）也是一大挑战。此外，聚类的效果常受数据分布影响，需较强的探索性数据分析能力。
> - **推荐人群**： 有兴趣探索无监督学习的同学，或希望提升数据分析和聚类技能的小组。

在GPT的分析基础上，我们各自对5个选题进行了2天左右的调研分析，并在线上进行了讨论。

DJ 11-13 20:17:22

感觉那种在线文本情感分析的网站的效果还挺有待改进的

5677 11-13 20:18:07

最常规的方法应该是SVM

DJ 11-13 20:18:50

先选一下调研的题目吧

DJ 11-13 20:22:18

那个房产的那个感觉也有点难。。

DJ 11-13 20:59:40

我想选第5个调研，无监督聚类的那个🤔

将芜 11-13 21:00:07

我先调研一下第4题吧

5677 11-13 21:05:49

那我看看第二题

Sollkatt 11-13 21:06:48

我看一下第一题

一笑奈何 11-13 21:07:03

我也看第五题吧

DJ 11-15 15:22:14

确实有案例，我感觉那个租房那个不太好，因为感觉特征工程不好做。之前我看别人做过。就是他。多说。不知道能产出什么结果。得不出什么结论

DJ 11-15 22:23:05

甯聿 邀请您参加腾讯会议
会议主题：甯聿的快速会议
会议时间：2024/11/15 22:22-23:22 (GMT+08:00) 中国标准时间 - 北京

点击链接直接加入会议：
https://meeting.tencent.com/dm/DP6ruNgLbl2Q

#腾讯会议：381-286-020

复制该信息，打开手机腾讯会议即可参与

5677 11-15 23:01:42

我们排除了四、五，觉得二不错

Sollkatt 11-18 18:30:56

关于分工我说一下我的想法吧，大家有什么看法也可以提

Sollkatt 11-18 18:31:12

1.探索性数据分析（1人），看一下整体以及各个类别数据的情况和特点，可以多画几个图放报告里，配上描述的文字，可以参考这个Modelling之前的部分 https://www.kaggle.com/code/georgesaavedra/best-sentiment-classifier-transformers#RoBERTa

Sollkatt 11-18 18:31:22

2.关键词提取（1人） 比如用TF-IDF算法把词转换成向量，助教

我们分析了各个选题的特点、需要的技术栈和我们的能力，最后选择了文本情感分析。虽然我们小组没有自然语言处理基础，但希望提升文本数据处理技能，因此选择了这个

选题。

2. 问题背景

> "There's a thin line between likably old-fashioned and fuddy-duddy, and The Count of Monte Cristo ... never quite settles on either side."“在讨人喜欢的老式风格和老顽固之间有一条细细的界线，而基督山伯爵……从未在任何一边站稳脚跟。”

这是一条电影评论，来自烂番茄电影评论数据集。它是一个用于情感分析的电影评论语料库，最初由 Pang 和 Lee [2] 收集。在他们对情感树库的工作中，Socher 等人 [3] 使用 Amazon 的 Mechanical Turk 为语料库中的所有解析短语创建精细标签。

Kaggle 正在为机器学习社区举办这次比赛，以用于娱乐和 练习。本次比赛提供了一个机会，可以在烂番茄数据集上对我们的情感分析想法进行基准测试。句子否定、讽刺、简洁、语言歧义等障碍使这项任务非常具有挑战性。

该任务需要处理大量的文本数据，并进行情感分类。处理文本的难点在于要解决语言的多义性和复杂的语境问题，还需要对模型进行优化才能达到较高准确度。使用预训练模型如 BERT 可以提高效果，但对资源和模型调优的要求较高。

3. 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：

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

此次任务是一个NLP领域中的文本情感分析任务，本质是一个多分类问题，但数据是文本，所以需要将文本数据转化数值数据，以便分类器（或其他模型）学习。

挑战

1. **文本复杂性**：
    - 评论可能包含成语或俚语。

    - 讽刺或挖苦。
    - 模棱两可的语言。
    - 否定句（如 “不好”）
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。

研究方法和工作流程

1. **数据预处理和EDA**： 分析数据集的特征和结构，处理缺失值、异常值。对数据初步分析，探索有无规律
    1. 删除不必要的字符（标点符号、HTML 标记）。
    2. 将评论分词，应用停止词去除和词干化/词素化。
    3. 探索性数据分析，包括数据分布、类别分布、句长分布等。
2. **词向量表示**： 根据任务性质、数据集特征和模型来选择合适的向量化方法
    1. **词袋模型（Bag of Words）**： 将文本转换为词频向量，每个词一个维度，词频作为值。
    2. **TF-IDF 向量化**： 词频-逆文档频率（TF-IDF）权重，权重越高，代表该词在该文档中越重要。
3. **模型训练和调优**： 尝试不同文本分类模型（如 LSTM、BERT 等）和参数调优，提升模型效果
    - **简单模型**：
        1. **XGBoost**： 作为。
        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。
        - **LSTM**:
4. **结果分析**： 用合适的评价指标评价指标（如准确率、F1 分数等）衡量模型效果，并进行可视化展示 由于这是一个多分类任务，无法直接用AUC和ROC来直观可视化，只能间接地将其转化为多个二分类问题但不够直观，所以使用准确率、精确度、召回率、F1-分数和混淆矩阵等指标来评估。
5. **心得与总结**： 分析实验过程中的挑战和收获。
    1. 我们对数据集的分析和处理有所了解，对文本分类有了一定的认识。
    2. 我们尝试了不同模型，并对模型效果进行了评估，效果理想。
    3. 我们尝试了不同特征提取方法，效果理想。
    4. 我们尝试了不同参数调优方法，效果理想。

二、数据分析处理

数据描述和检视

以下是对于数据集的描述：

> 该数据集由制表符分隔文件组成，其中包含来自Rotten Tomatoes数据集的短语。为了基准测试，保留了 train/test 拆分，但句子相对原始顺序，已重新排列。每个句子都已被 Stanford 解析器解析为许多短语。每个短语都有一个 PhraseId。每个句子都有一个 SentenceId。重复的短语（如短/常用词）在数据中仅包含一次。
>
> train.tsv 包含短语及其关联的情绪标签。我们还提供了一个 SentenceId，以便您可以跟踪哪些短语属于单个句子。
>
> test.tsv 仅包含短语。您必须为每个短语分配一个情绪标签。
>
> 情绪标签包括： 0 - negative 1 - somewhat negative 2 - neutral 3 - somewhat positive 4 - positive

数据insight

> <class 'pandas.core.frame.DataFrame'> RangeIndex: 156060 entries, 0 to 156059 Data columns (total 4 columns):
>
> Column Non-Null Count Dtype
>
> ---
>
> 0 PhraseId 156060 non-null int32 1 SentenceId 156060 non-null int32 2 Phrase 156060 non-null object 3 Sentiment 156060 non-null int32 dtypes: int32(3), object(1) memory usage: 3.0+ MB None

	PhraseId	SentenceId	Phrase	Sentiment
0	1	1	A series of escapades demonstrating the adage ...	1
1	2	1	A series of escapades demonstrating the adage ...	2
2	3	1	A series	2
3	4	1	A	2
4	5	1	series	2

> <class 'pandas.core.frame.DataFrame'> RangeIndex: 66292 entries, 0 to 66291 Data columns (total 3 columns):

> Column Non-Null Count Dtype
>
> ---
>
> 0 PhraseId 66292 non-null int32 1 SentenceId 66292 non-null int32 2 Phrase 66291 non-null object dtypes: int32(2), object(1) memory usage: 1.0+ MB None

	PhraseId	SentenceId	Phrase
0	156061	8545	An intermittently pleasing but mostly routine ...
1	156062	8545	An intermittently pleasing but mostly routine ...
2	156063	8545	An
3	156064	8545	intermittently pleasing but mostly routine effort
4	156065	8545	intermittently pleasing but mostly routine

为了处理的简便起见，我们在这一部分导入数据时在原始的tsv文件中把列名删除了，重新指定了列名并指定了每一列的数据类型，防止出现以外的类型转换问题。由之前的一些测试代码可以发现本数据的质量较好，没有缺失数据等，故不需要做这一步预处理。

这里可看到测试集有一个空缺值，后面在文本向量化和建模的时候会导致报错。`ValueError: np.nan is an invalid document, expected byte or unicode string.` 一开始想用众数填充或其他填充方法，但想到原论文分词的逻辑并不会出现空词，于是特意看了下数据。

	PhraseId	SentenceId	Phrase
15516	171577	9213	None of this violates the letter of Behan 's b...
15517	171578	9213	None of this violates the letter of Behan 's b...
15518	171579	9213	None of this
15519	171580	9213	NaN
15520	171581	9213	violates the letter of Behan 's book , but mis...
15521	171582	9213	violates

根据分词逻辑，推测test分词时Sentence 9213 Phrase 171580的字面值为"None"，导致读入时误识别为了NoneType。这里将其还原为 "None" 字符串即可。（实测发现读取时指定dtype无法避免这个错误，故特殊处理）

数据预处理和EDA

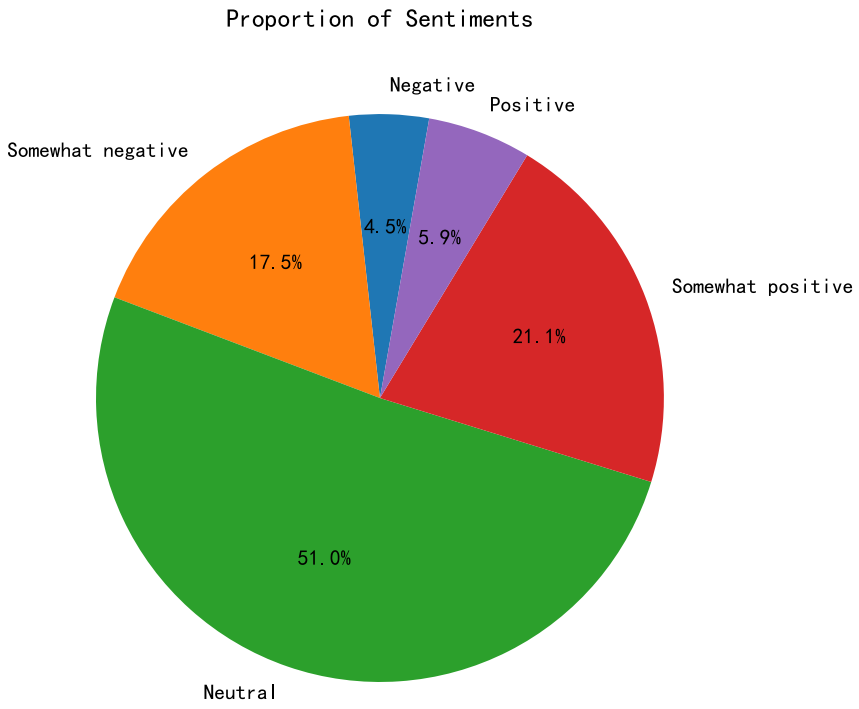
将评论分词。

应用停止词去除和词干化/词素化。

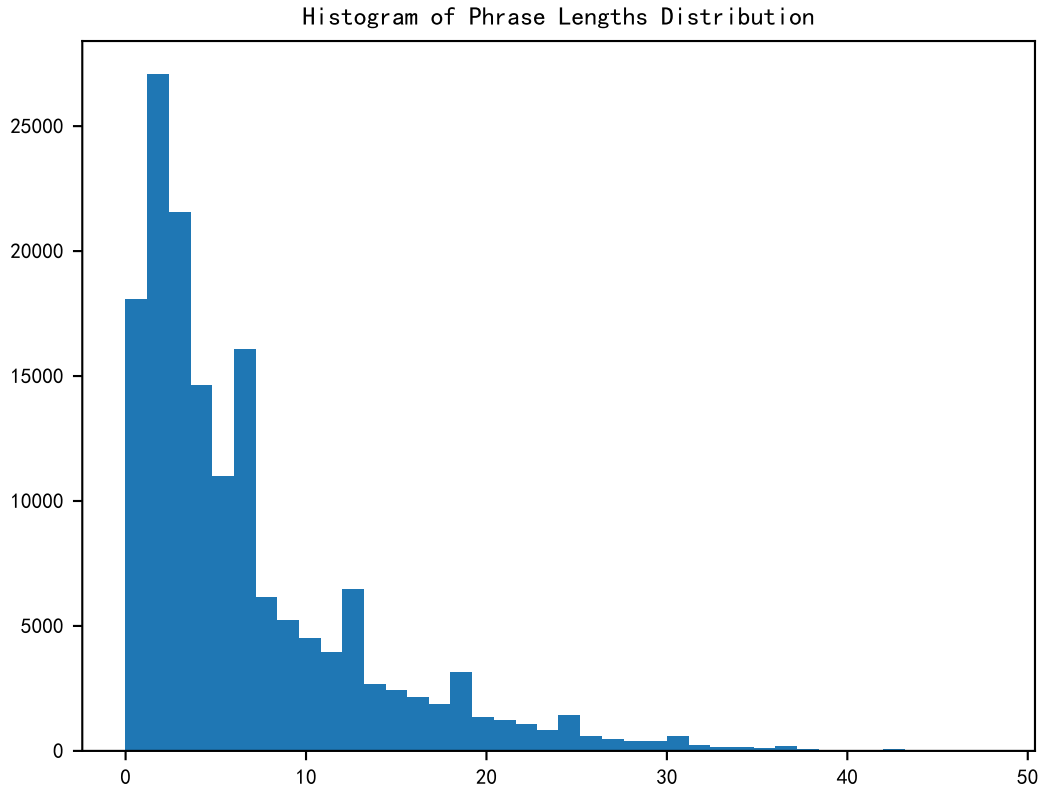
使用 TF-IDF 将文本转换为数字特征。

删除不必要的字符（标点符号、HTML 标记）。

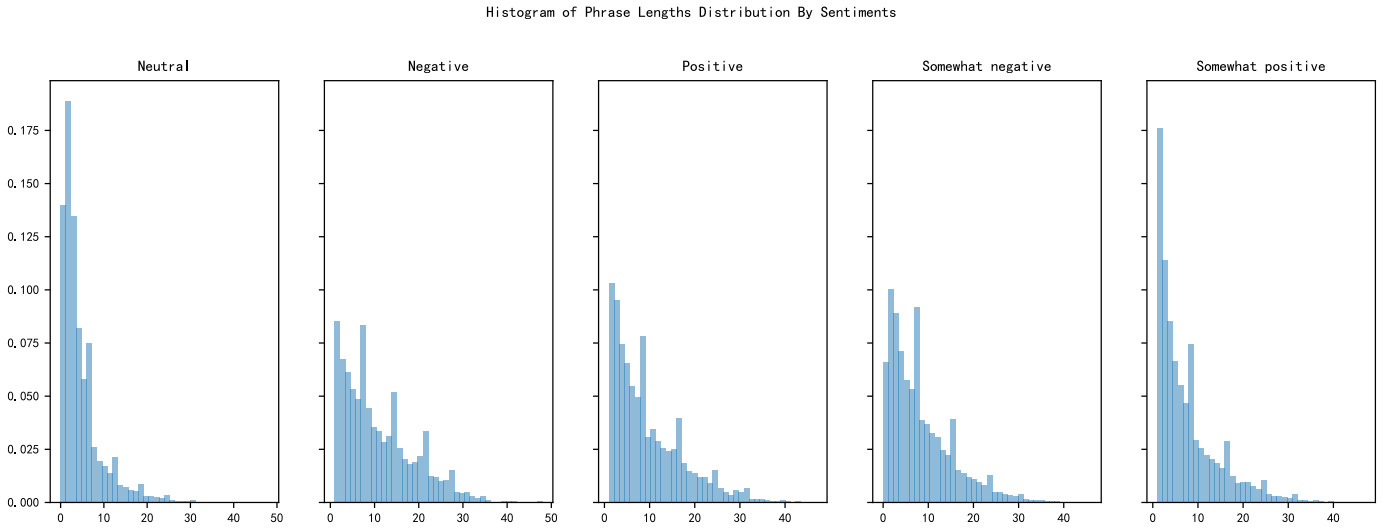
我们在这一步做了文本数据的预处理和一个初步统计。为了消除常见的干扰信息，我们去除了文本中的标点符号并将所有字母转为小写字母。我们对于情感类型的分布做了统计，结果显示Neutral占了一半以上，而最确定的Negative和Positive占比最少。这也非常符合本组数据将一个句子做树状拆分后，大部分短语的情感色彩为Neutral的特点。



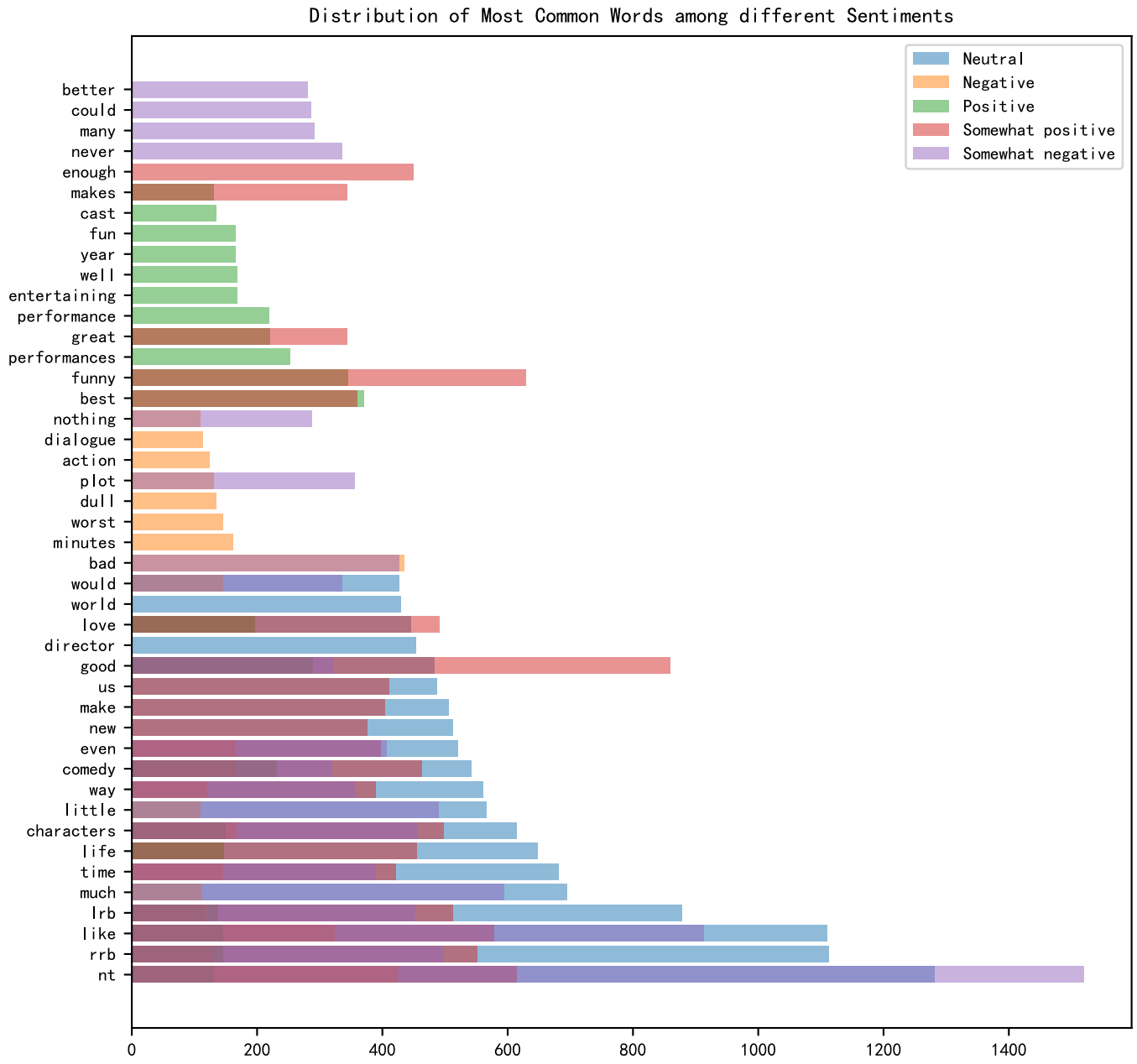
然后我们分析了总体句长的分布。本组数据的句长分布呈现典型的偏态分布特征。结合数据的来源，可以推测其相对符合指数分布。



然后我们对于不同情感色彩的句长进行了分析，探讨其差异性。可以看到Neutral组的句长分布明显偏短，而注释有情感的短句长度会更高一些。但是所有组别的分布仍然都是短句远多于长句。这一方面说明了大部分短句可能都没有明显的感情色彩，另一方面也说明了长句的感情色彩仍然需要由决定性的短语来确定。



随后我们利用 `nltk`进行了一个初步的词频分析.选取了每组词频前20位的词语，并在常规禁用词外增加了我们前期看到的一些在本任务中比较常见但没有很大价值的词汇。可以看到特别高频出现的词汇依旧没有太大的感情色彩，在各组之间均有分布。但是次高频出现的词汇就能体现出比较明显的感情色彩。这提示这些次高频词汇可能是分析文本感情的关键。



三、文本向量化

词袋模型

将文本转换为词频向量，每个词一个维度，词频作为值。

Countvectorizer只会对字符长度不小于2的单词进行处理，如果单词就一个字符，这个单词就会被忽略。 注意 ，经过训练后，CountVectorizer就可以对测试集文件进行向量化了，但是向量化出来的特征只是训练集出现的单词特征，如果测试集出现了训练集中没有的单词，就无法在词袋模型中体现了。

TF-IDF

词频-逆文档频率（TF-IDF）权重，权重越高，代表该词在该文档中越重要。scikit-learn库中的tf-idf转换与标准公式稍微不同，而且tf-idf结果会用L1或L2范数进行标准化。

四、模型训练和调优

实施细节

模型 1：xgBoost

- 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：LSTM

- 特征提取： TF-IDF 向量器。
- 优势：。
- 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。

五、实验结果分析

模型性能

指标	朴素贝叶斯	随机森林	BERT	LSTM
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. 朴素贝叶斯提供了一个快速、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。

六、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。

难点

微调 BERT 的计算成本。

处理文本中的边缘情况，如讽刺和模棱两可的表达。

未来的工作

1. 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。
2. 处理文本中的边缘情况，如讽刺和模棱两可的表达。

七、结论与建议

结论

- 朴素贝叶斯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

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	BERT
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

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. Will Cukierski. Sentiment Analysis on Movie Reviews. https://kaggle.com/competitions/sentiment-analysis-on-movie-reviews, 2014. Kaggle.
2. [2] Pang and L. Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In ACL, pages 115–124.
3. [3] Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank, Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Chris Manning, Andrew Ng and Chris Potts. Conference on Empirical Methods in Natural Language Processing (EMNLP 2013).
4. 【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