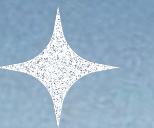


MBTI 人格测试

基于机器学习的分类问题



中国人民大学统计学院

卢虹臻、王梓尧、刘亚林、唐子炫、吕淞彬

目录

引言

数据收集与预处理

方法和实验设计

实验结果和分析

结论

引言

[返回目录页面](#)

MBTI介绍

MBTI (Myers-Briggs Type Indicator) 是一种广泛使用的人格类型测量工具，旨在帮助人们了解自己的个性特征和倾向。

MBTI的核心概念是人格类型，它将个体的行为和心理偏好归类为四对相对维度。每个维度包括两个极端特质，人们倾向于在每个维度上更偏好其中一个特质。以下是MBTI的四对维度：

- 性格取向 (Attitudes)：
 - 外向 (Extraversion) 与内向 (Introversion)
- 感知方式 (Perceiving)：
 - 感觉 (Sensing) 与直觉 (Intuition)
- 决策方式 (Judging)：
 - 思考 (Thinking) 与情感 (Feeling)
- 生活方式 (Lifestyle)：
 - 判断 (Judging) 与知觉 (Perceiving)

评论文本挖掘涉及多种方法和技术，包括自然语言处理（NLP）、文本分类、情感分析、主题建模等。这使得我们可以充分使用本学期所学的内容，使用多种机器学习方法自动化地处理和分析大量的评论文本数据。

1. 自然语言处理：NLP技术用于对评论文本进行文本清洗、分词、词性标注等预处理步骤，以便进行后续的分析和建模。
2. 文本分类：文本分类技术通过训练机器学习模型，将评论文本分为不同的类别或情感极性，如正面、负面或中性评论。这有助于快速了解大规模评论数据的整体倾向。
3. 情感分析：情感分析技术用于识别评论文本中的情感倾向，如喜欢、讨厌、满意或失望等。这对于了解用户的情感反应至关重要。
4. 深度学习方法：探索深度学习技术在评论文本挖掘中的应用，以提高模型的准确性和效果。

文本挖掘意义

数据收集与预处理

[返回目录页面](#)

数据集介绍

数据集是基于Kaggle网站的MBTI人格类型Twitter数据集，该数据集旨在探索Twitter用户的推文内容与其自报的MBTI人格类型之间的关联。数据集包含了来自Twitter的用户推文数据以及用户自报的MBTI人格类型，样本量为8675。

```
1 # 发现仅有两个变量，且数据无缺失值
2 data.info()
```

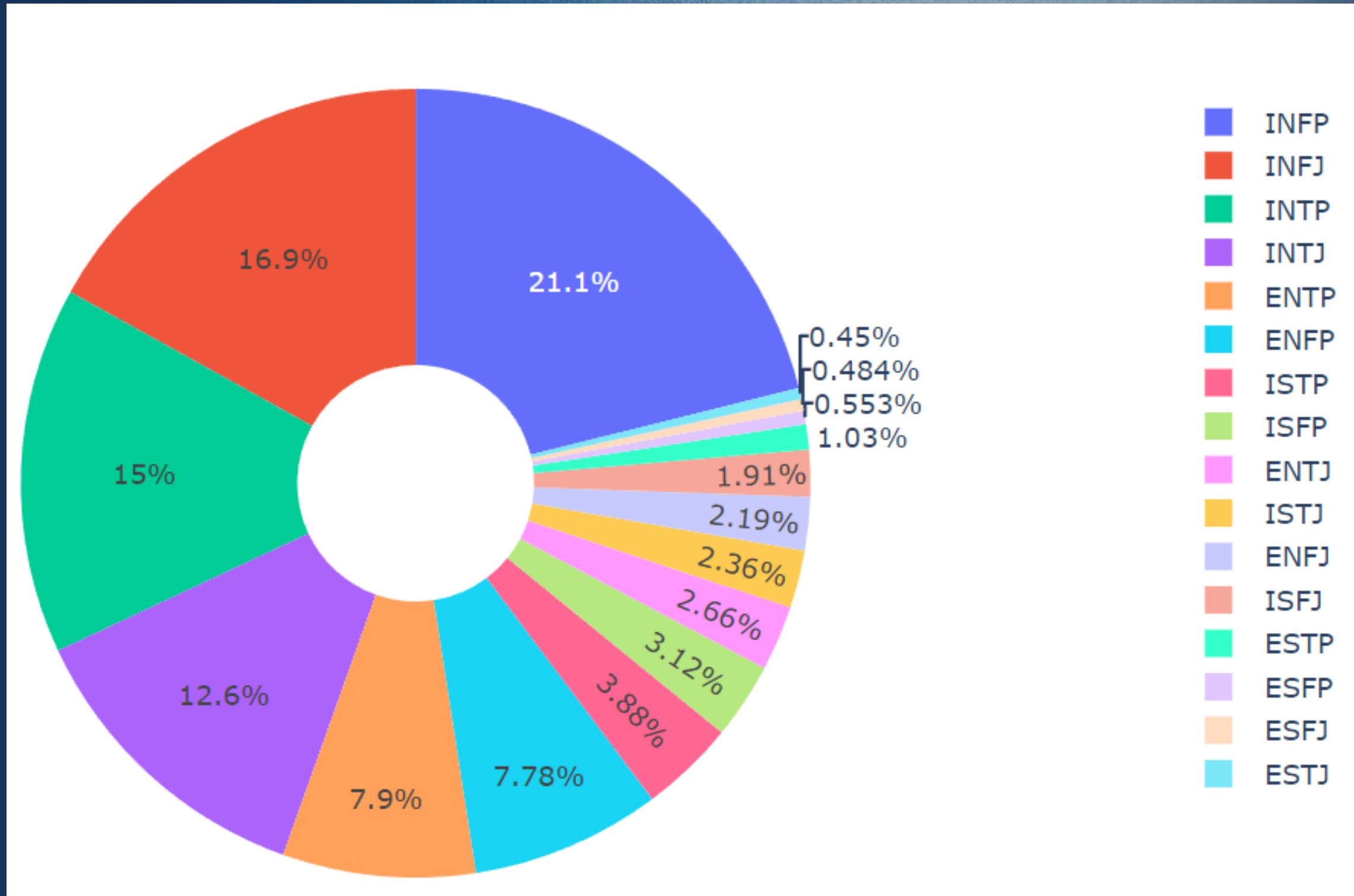
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8675 entries, 0 to 8674
Data columns (total 2 columns):
 #   Column  Non-Null Count  Dtype  
---  --  
 0   type    8675 non-null   object 
 1   posts   8675 non-null   object 
dtypes: object(2)
memory usage: 135.7+ KB
```

```
1 # 读入数据
2 data=pd.read_csv('mbti_1.csv')
3 data.head(10)
```

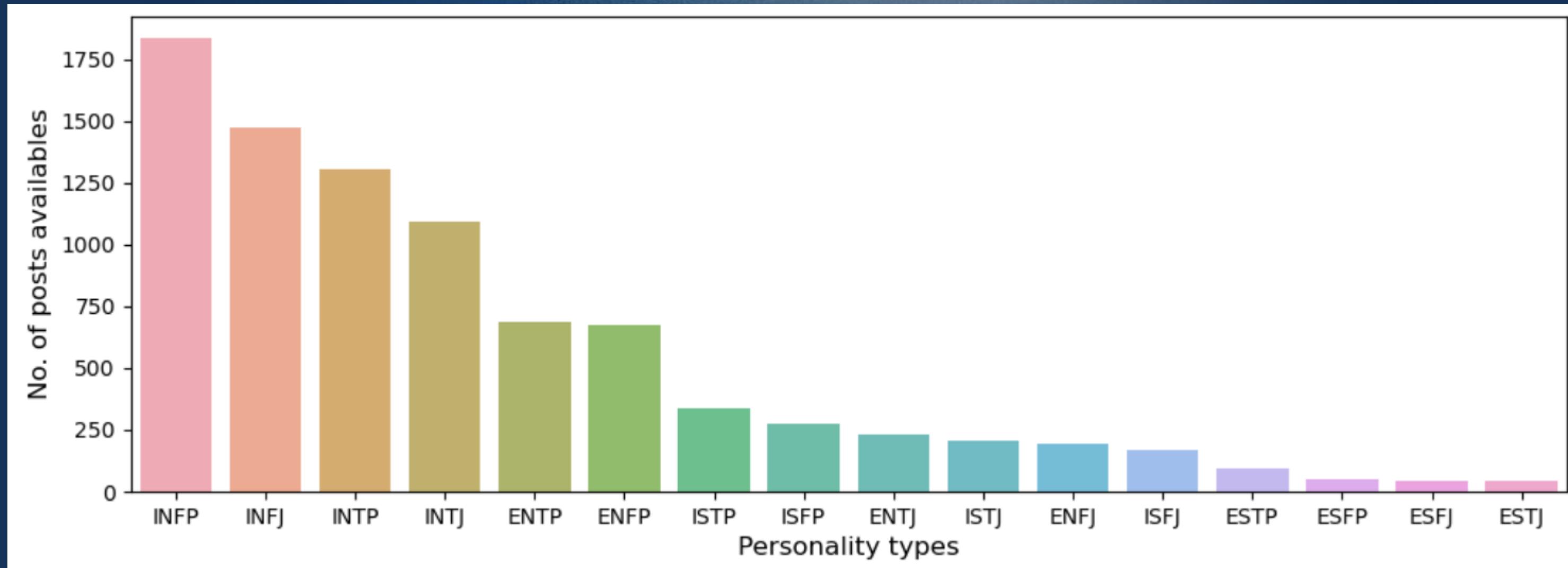
| | type | posts |
|---|------|---|
| 0 | INFJ | 'http://www.youtube.com/watch?v=qsXHcwe3krw ... |
| 1 | ENTP | 'I'm finding the lack of me in these posts ver... |
| 2 | INTP | 'Good one ____ https://www.youtube.com/wat... |
| 3 | INTJ | 'Dear INTP, I enjoyed our conversation the o... |
| 4 | ENTJ | 'You're fired. That's another silly misconce... |
| 5 | INTJ | '18/37 @.@@ Science is not perfect. No scien... |
| 6 | INFJ | 'No, I can't draw on my own nails (haha). Thos... |
| 7 | INTJ | 'I tend to build up a collection of things on ... |
| 8 | INFJ | I'm not sure, that's a good question. The dist... |
| 9 | INTP | 'https://www.youtube.com/watch?v=w8-egj0y8Qs ... |

数据可视化

各人格类型占比饼状图



数据可视化

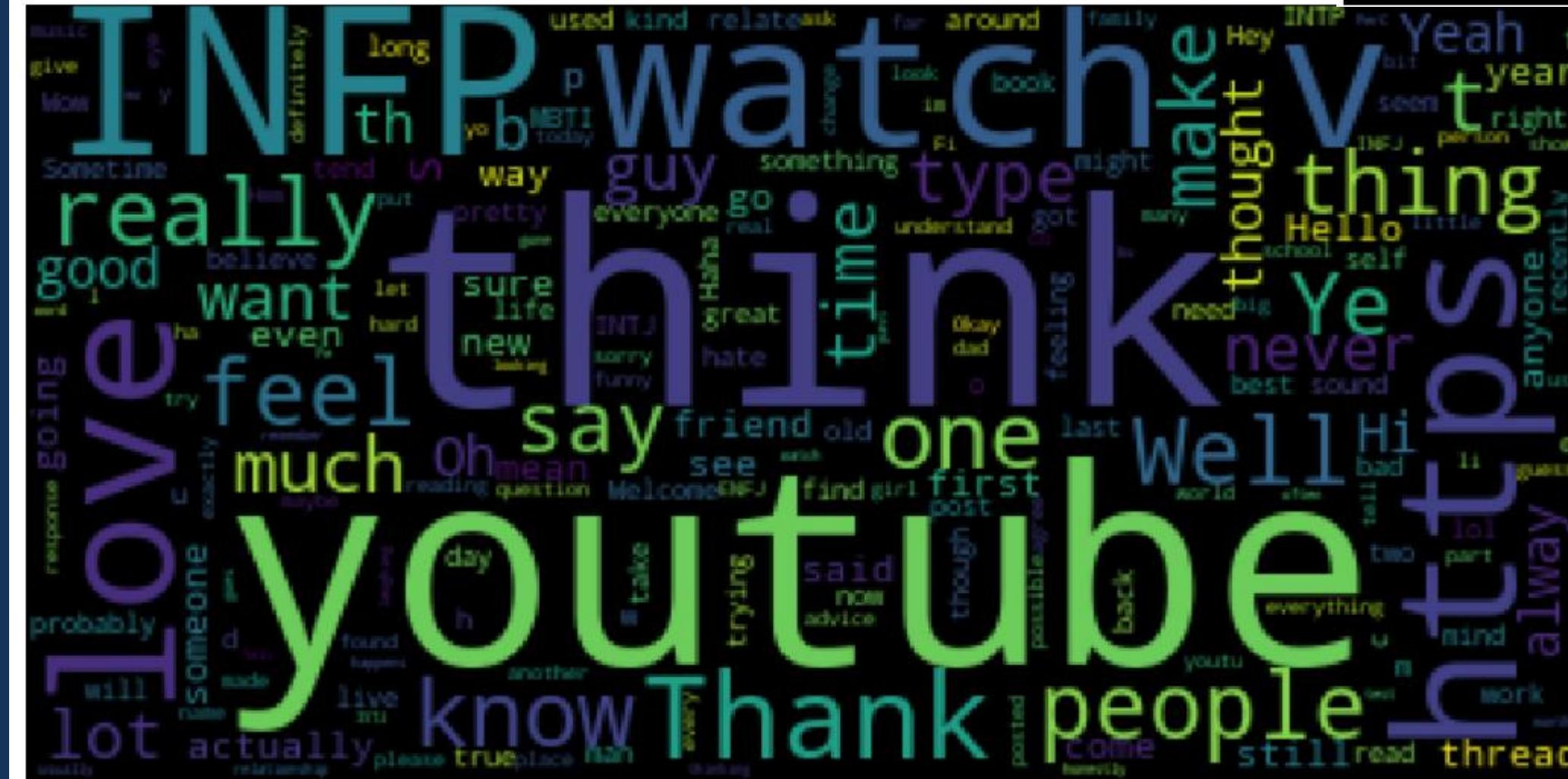


各人格类型数量条状图

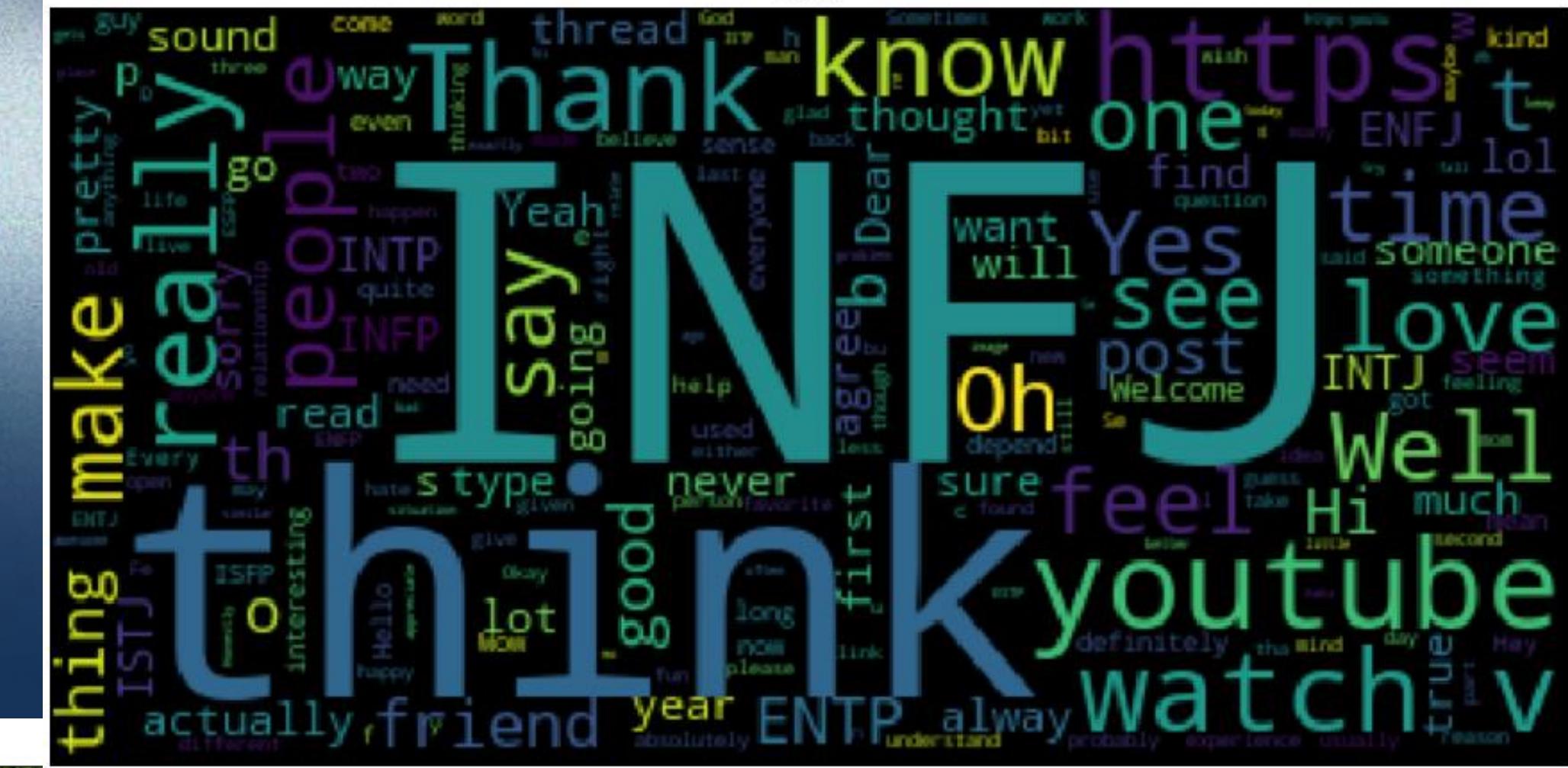
数据可视化

词云图：INFP

INFO



INR



词云图：INFJ

文本清洗

大小写转换

将评论文本转换为小写形式，以避免大小写的差异对情感分析结果的影响。

表情转换

使用EMOJI库将评论中的表情符号转换为对应的语义表达

去除非英文字符

使用正则表达式去除评论中的非英文和非数字的符号，以确保只保留有意义的文本内容。

词性还原和分词：

使用NLTK库中的WordNetLemmatizer和word_tokenize函数，对评论进行词性还原和分词操作，以便获得单词的基本形式和分词结果。

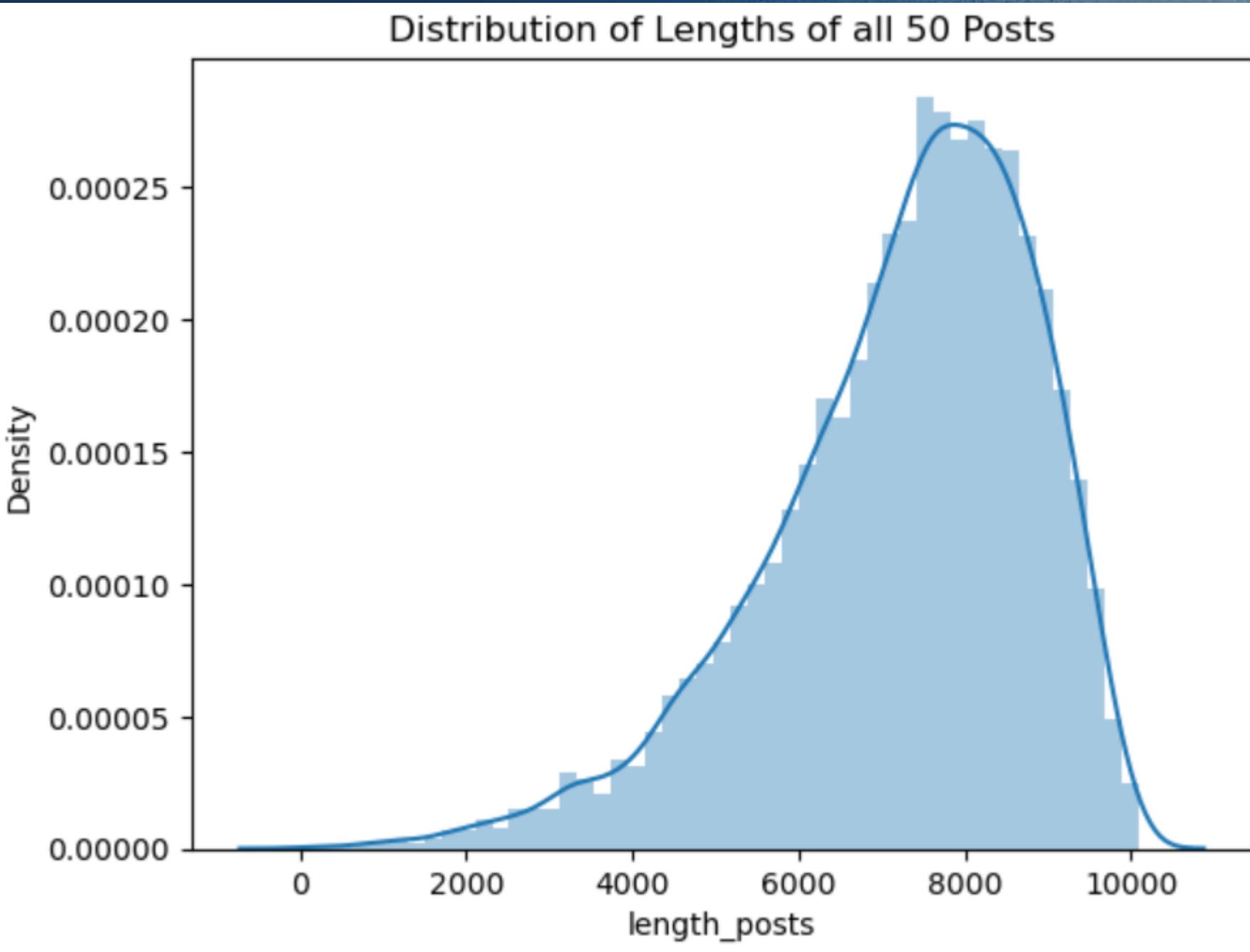
去除网址

使用正则表达式去除评论中的网址，因为网址对情感分析任务没有实际意义。

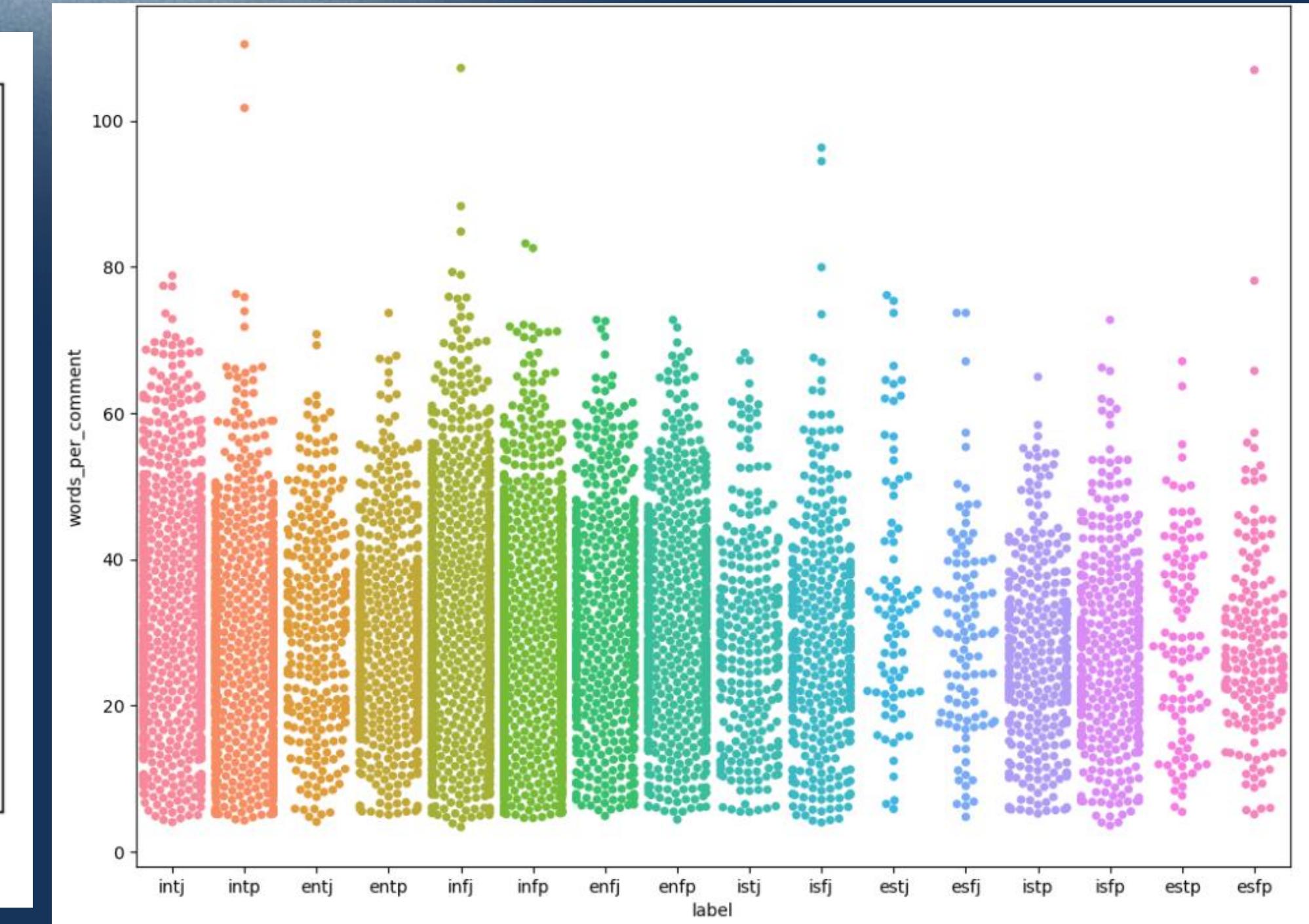
去除停用词

使用NLTK库中的停用词列表，去除评论中的停用词，以减少对情感分析结果的干扰。

数据可视化

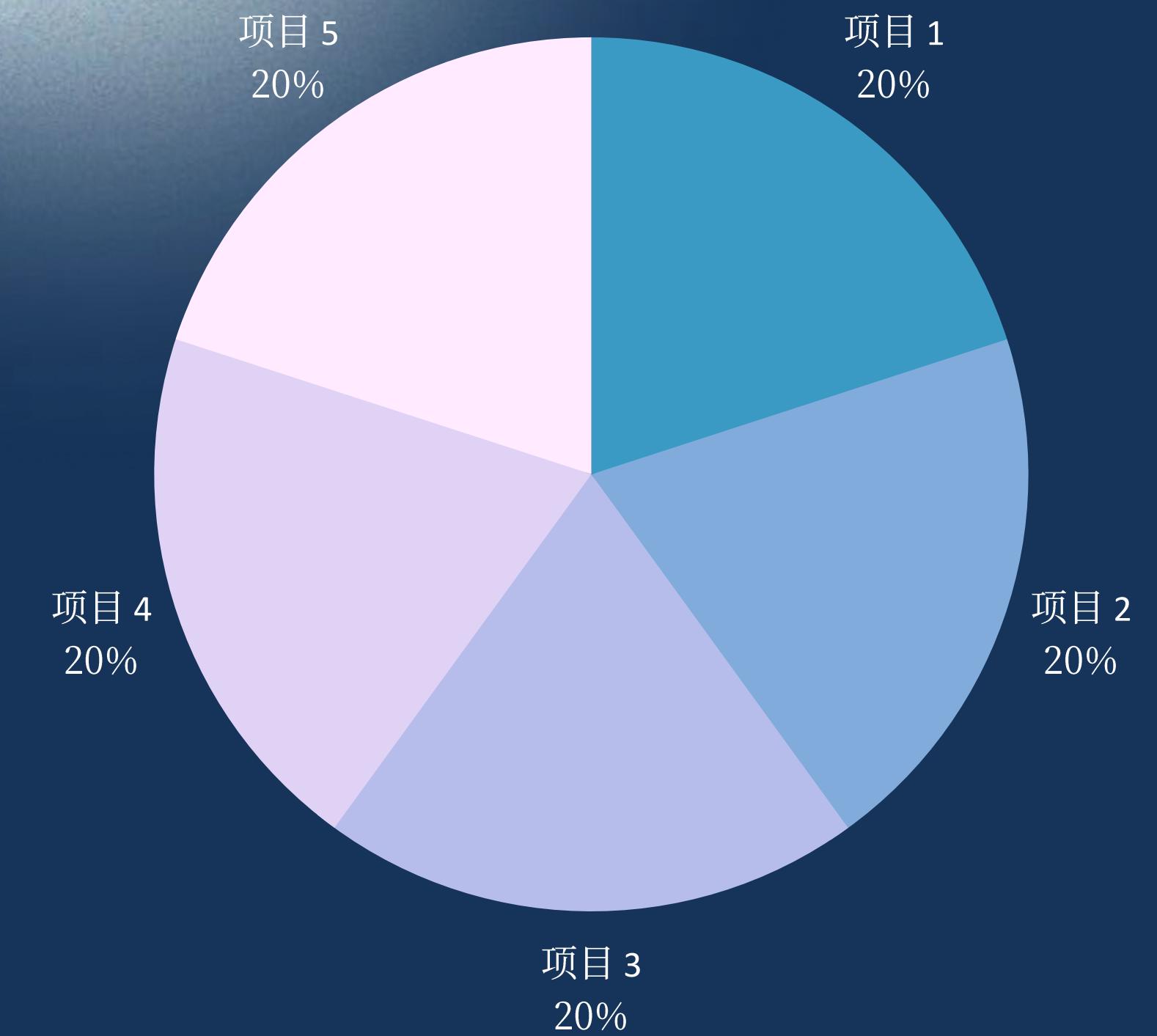


文本总长度分布图

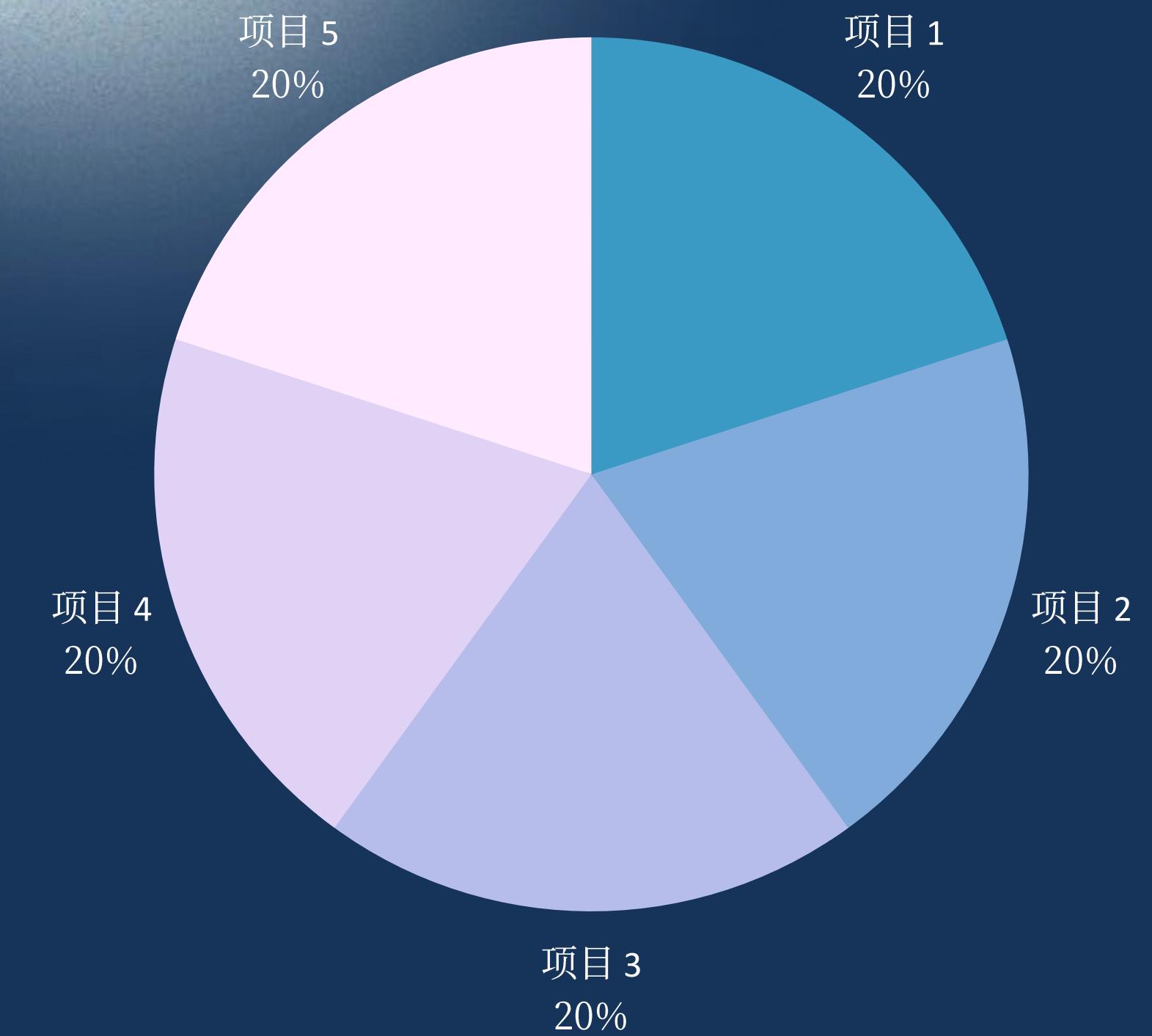


单条文本平均长度分布图

数据可视化



数据可视化



文本词编码过程

词向量WORD2VEC编码

在Word2vec模型中，根据词的上下文我们采用skip-gram算法进行编码训练，采用5作为一个句子中当前单词和被预测单词的最大距离。对于数据集中7811条评论数据，采用每条评论中的前400个词语作为句子特征进行句子编码，如果评论词数少于400词则用0填充，最终得到7811x400维的训练集向量。

TF-IDF编码

TF-IDf通过统计词语在文档中的出现次数和在整个文档集合中的分布情况来计算这些权重，并使用这些权重来过滤掉常见的词语，保留重要的词语。为对其进行的描述性分析中，我们将最大特征值数量设置为1000（否则电脑运行不出后面的描述性分析），并发现提取特征后特征词表有许多明显与人的性格和情感特征有关的词语。

词袋BOW编码

BOW的工作原理基于词频，它的基本思想是对于一个文本，忽略其词序和语法、句法，仅将其看做是一些词汇的集合。将所有文本的词放在一起得到语料库，汇编成词典，再将每个条文本用词典中每个单词出现的次数表示出来。在训练过程中，通过计算向量的余弦距离来计算两个文本间的相似度。在实际处理的过程中，还会进行词频统计这一步，只选择其中出现频率最高的前5000个词作为最终的词表。

方法和实验设计

模型优势



朴素贝叶斯

选择朴素贝叶斯的原因

1. 朴素贝叶斯模型发源于古典数学理论，有稳定的分类效率。

2. 对缺失数据不太敏感，算法也比较简单，常用于文本分类。

3. 分类准确度较高，不需要调参，速度快。

KNN

1. K-近邻法的模型复杂度更高（K较小时），更适合解决非线性分类问题。

2. K-近邻法考虑样本之间的相似性，这与Word2vec模型的特点相合，可以很好的考虑词语的情感信息。

3. K-近邻法是一种基于局部的学习，无数据输入设定。

ADABOOST

1. 集成学习可以解决预测模型（如决策树）的高方差问题。

2. 集成学习将一组弱模型联合起来使其成为一个强模型，提高预测性能。

3. 作为简单的二元分类器时，构造简单，不需要做特征筛选，结果可解释。

4. 在Adaboost的框架下，可以使用各种分类模型来构建弱学习器，非常灵活。

逻辑斯特回归

1. 逻辑斯特回归是一种广义线性模型，可以直接得到预测结果的概率值，便于评估置信度和风险。

2. 逻辑斯特回归可以处理稀疏的文本特征矩阵，不需要进行特征转换或归一化。

3. 逻辑斯特回归可以解释每个特征对输出变量的影响程度，有助于分析文本数据的重要信息和关键词。

决策树

- 1.TF-IDF编码将生成大量的特征向量等其他编码形式同样如此，而决策树模型可以较好地处理高维度的特征空间。
- 2.TF-IDF编码将赋予不同的特征不同的权重，而决策树模型可以自动选择重要的特征，方便地进行剪枝、合并、平滑等操作，以防止过拟合或欠拟合，减少冗余和无关的特征对预测的影响。

随机森林

- 1.随机森林模型在构建每个决策树时，会随机选择部分样本和部分特征，增加了样本和特征的多样性，减少了过拟合的风险。
- 2.随机森林模型可以输出每个类别的概率估计，以及每个特征的重要性评分，有助于分析文本数据的分类结果和特征贡献。
- 3.随机森林模型可以并行地训练多个决策树，提高了计算效率和速度。

支持向量机

- 1.支持向量分类机是有很好的理论支撑同时实际效果很好的分类算法。
- 2.支持向量分类机在不同核函数下能应对不同的问题。
- 3.支持向量分类机在进行高维的分类预测问题上有很多优势。

神经网络

使用Hugging Face网站中开源的BERT模型“paraphrase-MiniLM-L6-v2”作为神经网络的嵌入层（embedding layer）将用户的前100条评论转换成对应的句向量，组成维度为（100， 384）的评论矩阵，依据TextCNN模型搭建神经网络，使用准确率（accuracy）作为模型的评估指标。

- 通过多个二分类模型实现多目标分类任务
- 多目标分类模型

实验结果和分析

[返回目录页面](#)

朴素贝叶斯

从上到下，从左到右依次为 I/E, N/S, F/T, P/J的测试集评价结果，从结果可以看出，朴素贝叶斯的分类效果较为良好，其平均准确率达到了0.647。

朴素贝叶斯分类器的测试误差:0.326

测试样本的评价结果:

| | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| e | 0.00 | 0.00 | 0.00 | 764 |
| i | 0.67 | 1.00 | 0.81 | 1580 |
| accuracy | | | 0.67 | 2344 |
| macro avg | 0.34 | 0.50 | 0.40 | 2344 |
| weighted avg | 0.45 | 0.67 | 0.54 | 2344 |

朴素贝叶斯分类器的测试误差:0.230

测试样本的评价结果:

| | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| n | 0.77 | 1.00 | 0.87 | 1805 |
| s | 0.00 | 0.00 | 0.00 | 539 |
| accuracy | | | 0.77 | 2344 |
| macro avg | 0.39 | 0.50 | 0.44 | 2344 |
| weighted avg | 0.59 | 0.77 | 0.67 | 2344 |

朴素贝叶斯分类器的测试误差:0.404

测试样本的评价结果:

| | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| f | 0.60 | 0.99 | 0.75 | 1395 |
| t | 0.56 | 0.01 | 0.03 | 949 |
| accuracy | | | 0.60 | 2344 |
| macro avg | 0.58 | 0.50 | 0.39 | 2344 |
| weighted avg | 0.58 | 0.60 | 0.46 | 2344 |

朴素贝叶斯分类器的测试误差:0.439

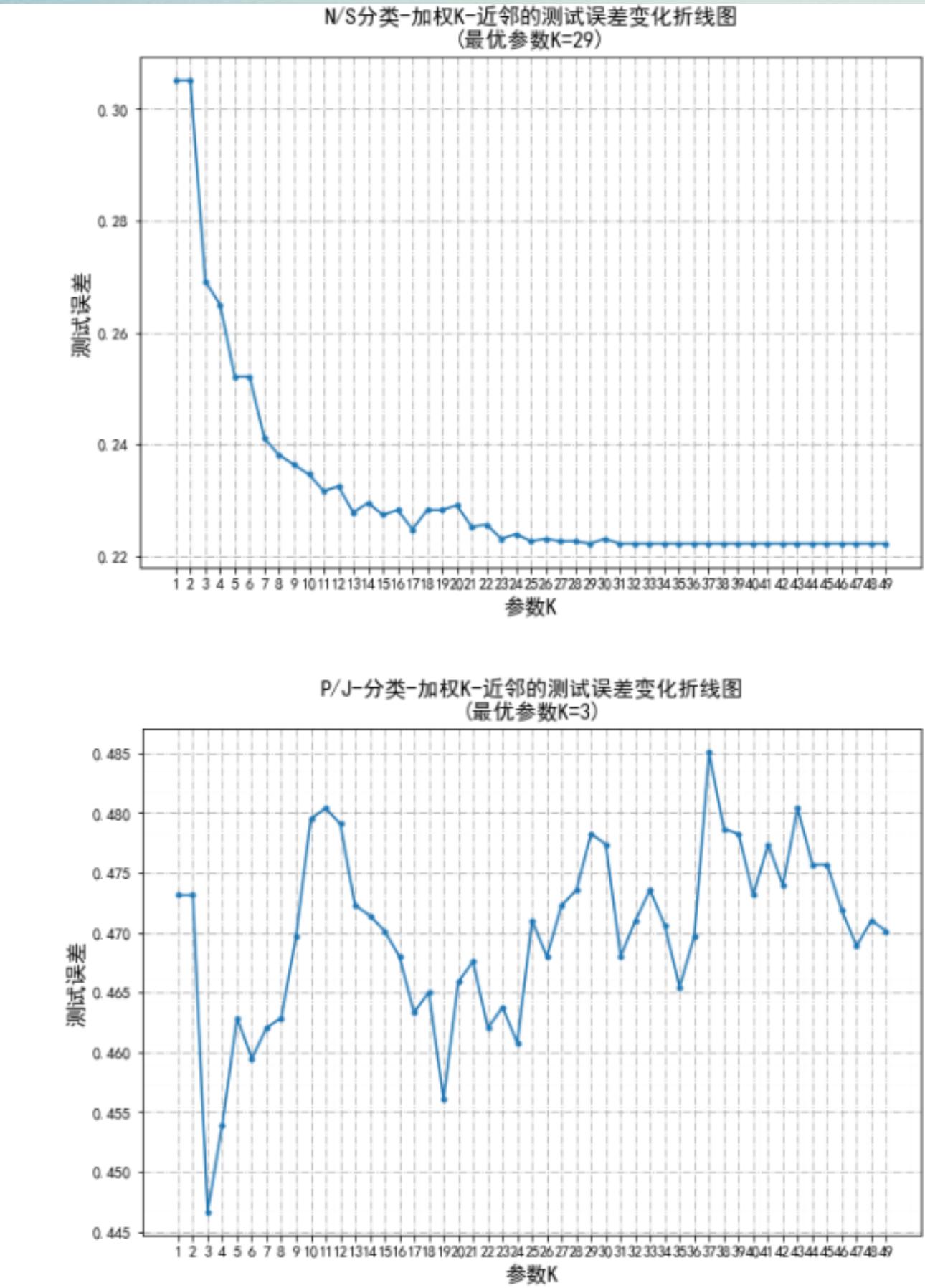
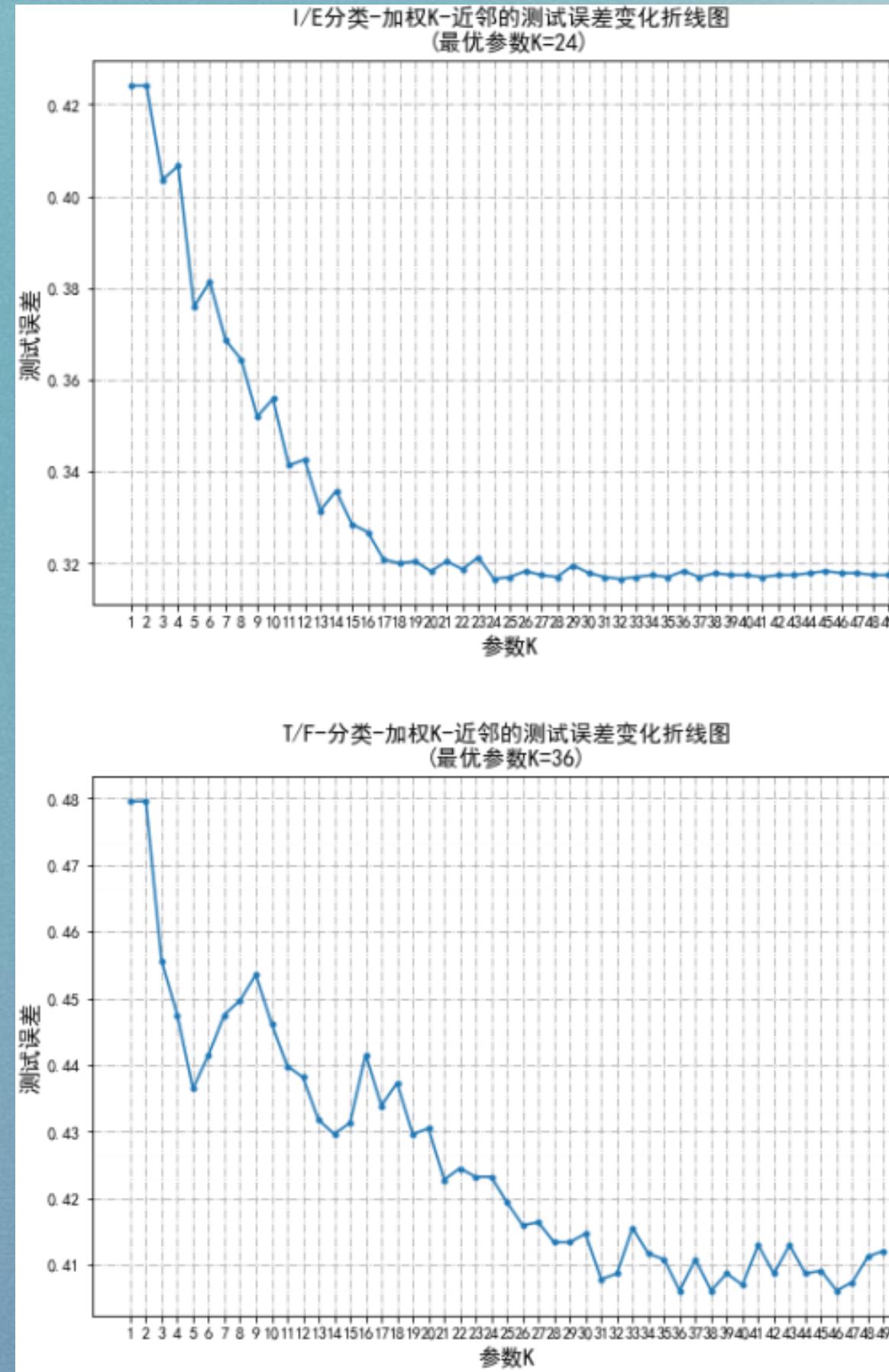
测试样本的评价结果:

| | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| j | 0.00 | 0.00 | 0.00 | 1027 |
| p | 0.56 | 1.00 | 0.72 | 1317 |
| accuracy | | | 0.56 | 2344 |
| macro avg | 0.28 | 0.50 | 0.36 | 2344 |
| weighted avg | 0.32 | 0.56 | 0.40 | 2344 |

朴素贝叶斯分类器判断出了几乎所有的多数类样本，但放弃了全部的少数类样本，很容易受到样本不均衡问题影响

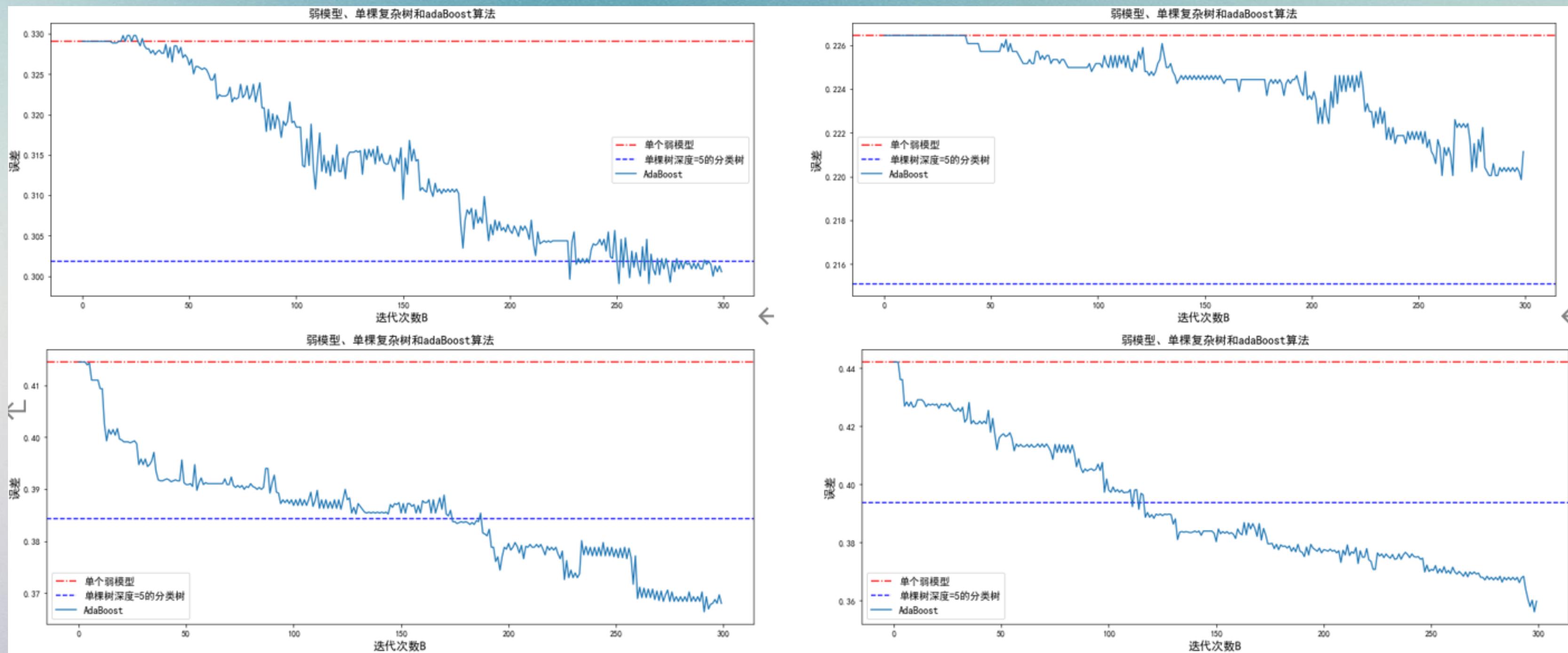
KNN

从上到下，从左到右依次为 I/E, N/S, F/T, P/J的最佳K近邻值，可以看到前三个维度的分类器的误差都大致随着K的增大而逐渐下降，而P/J的最佳K值为3且误差较大。综合来看，KNN模型的分类预测效果较为良好，在训练集上的的平均准确率达到了0.652。



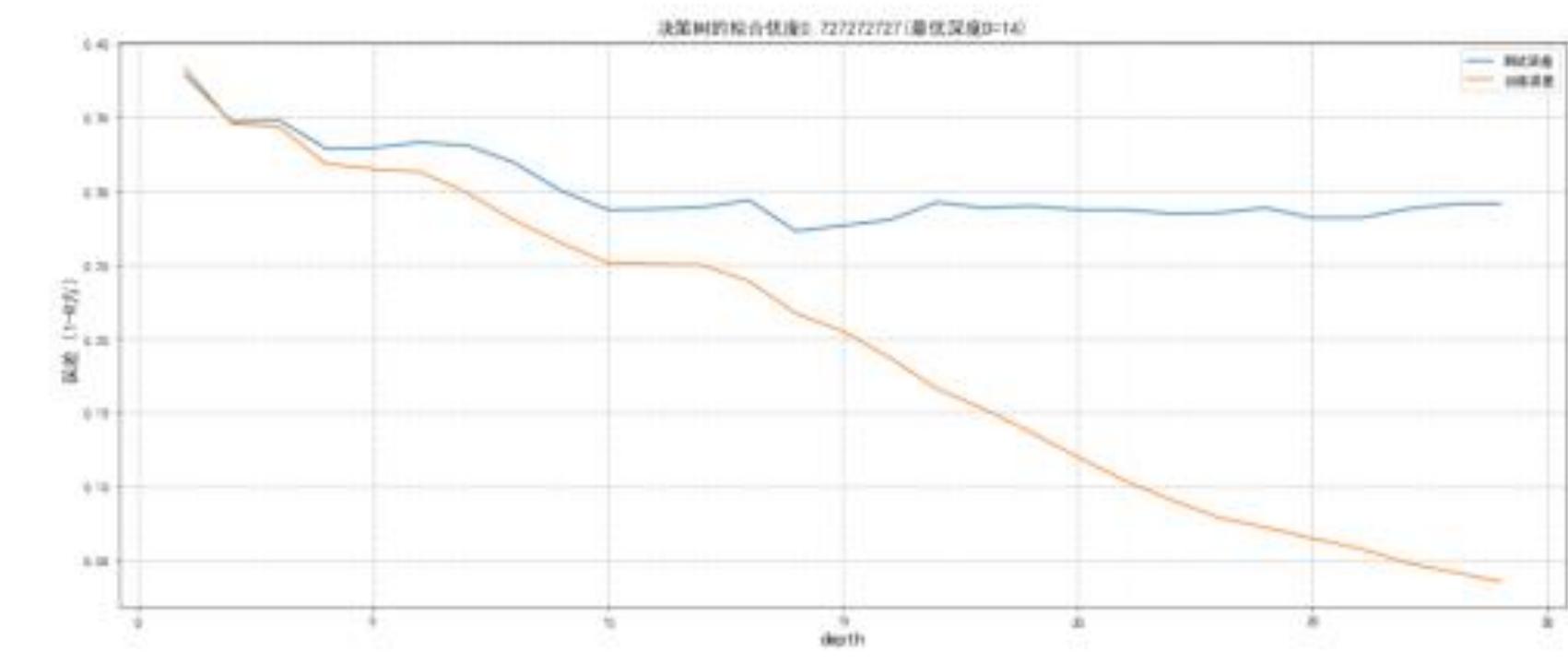
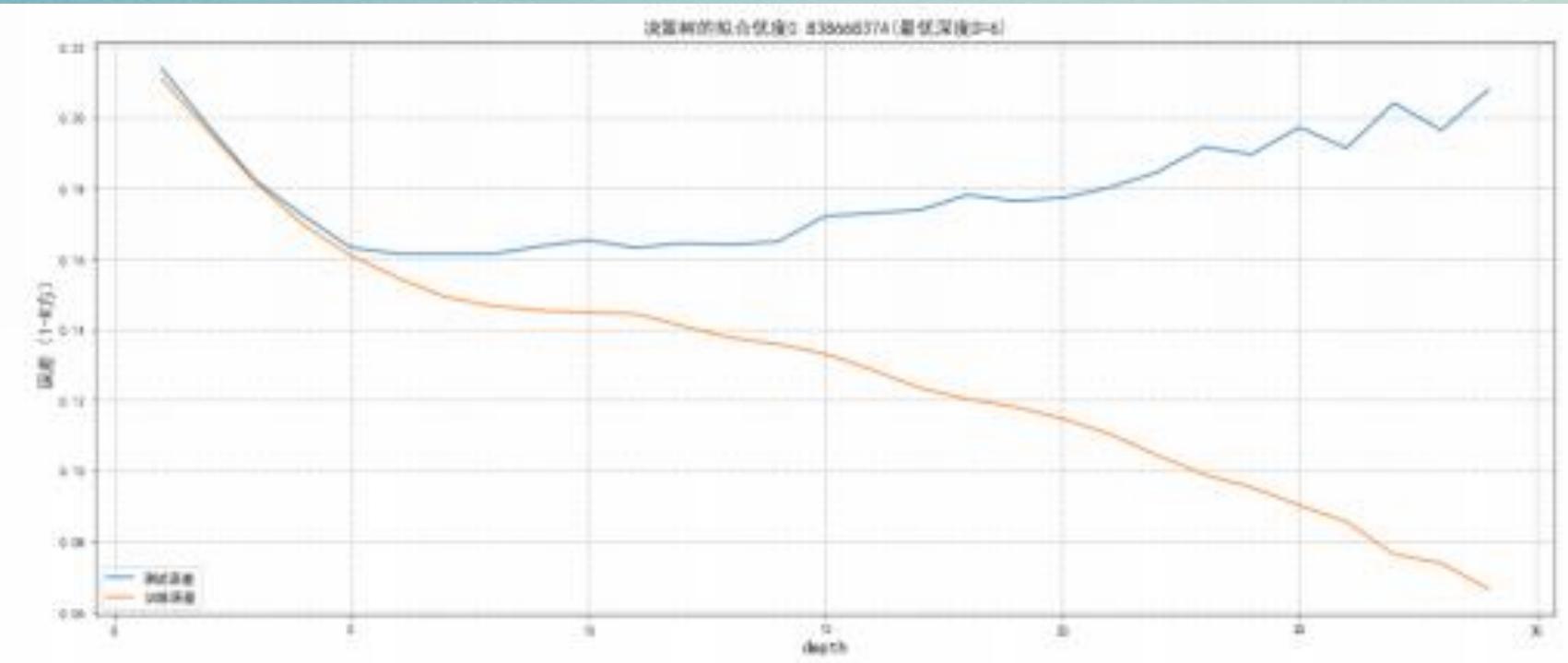
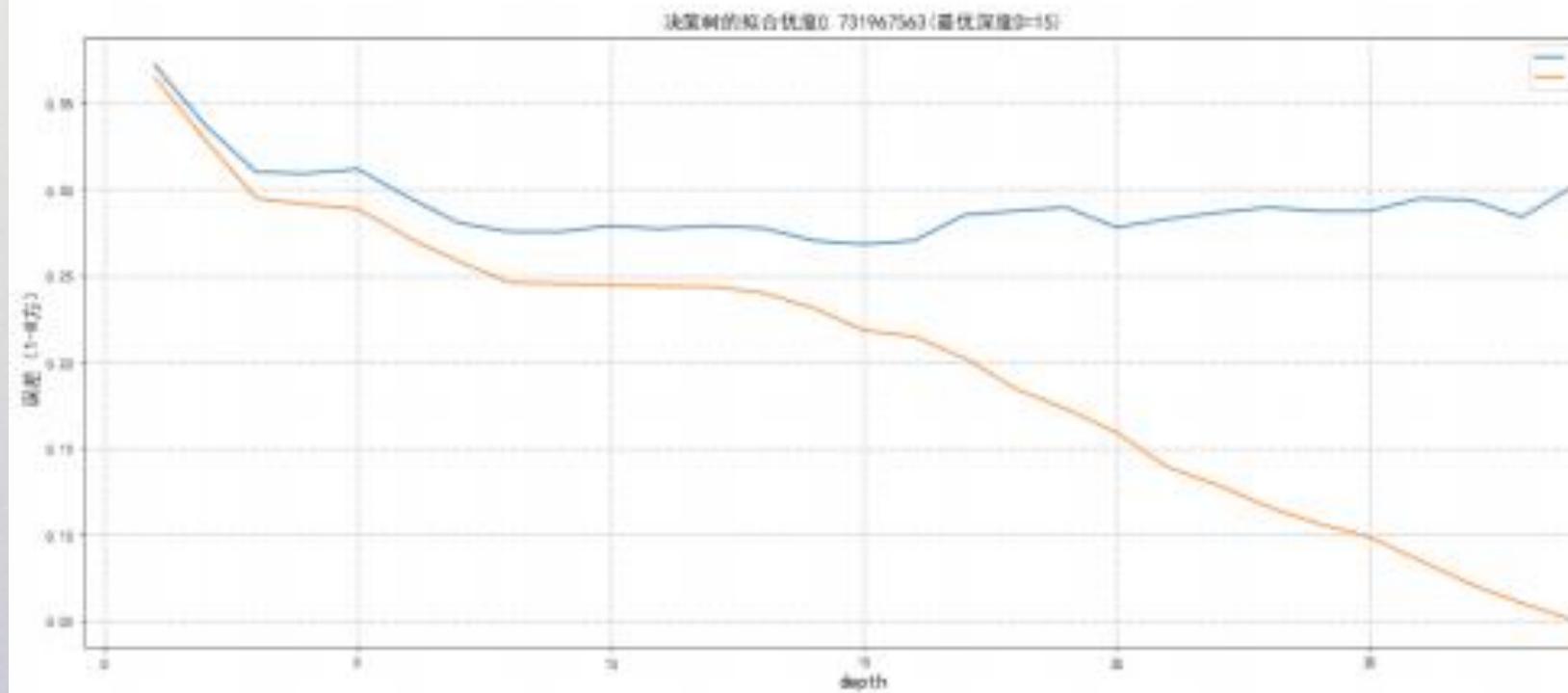
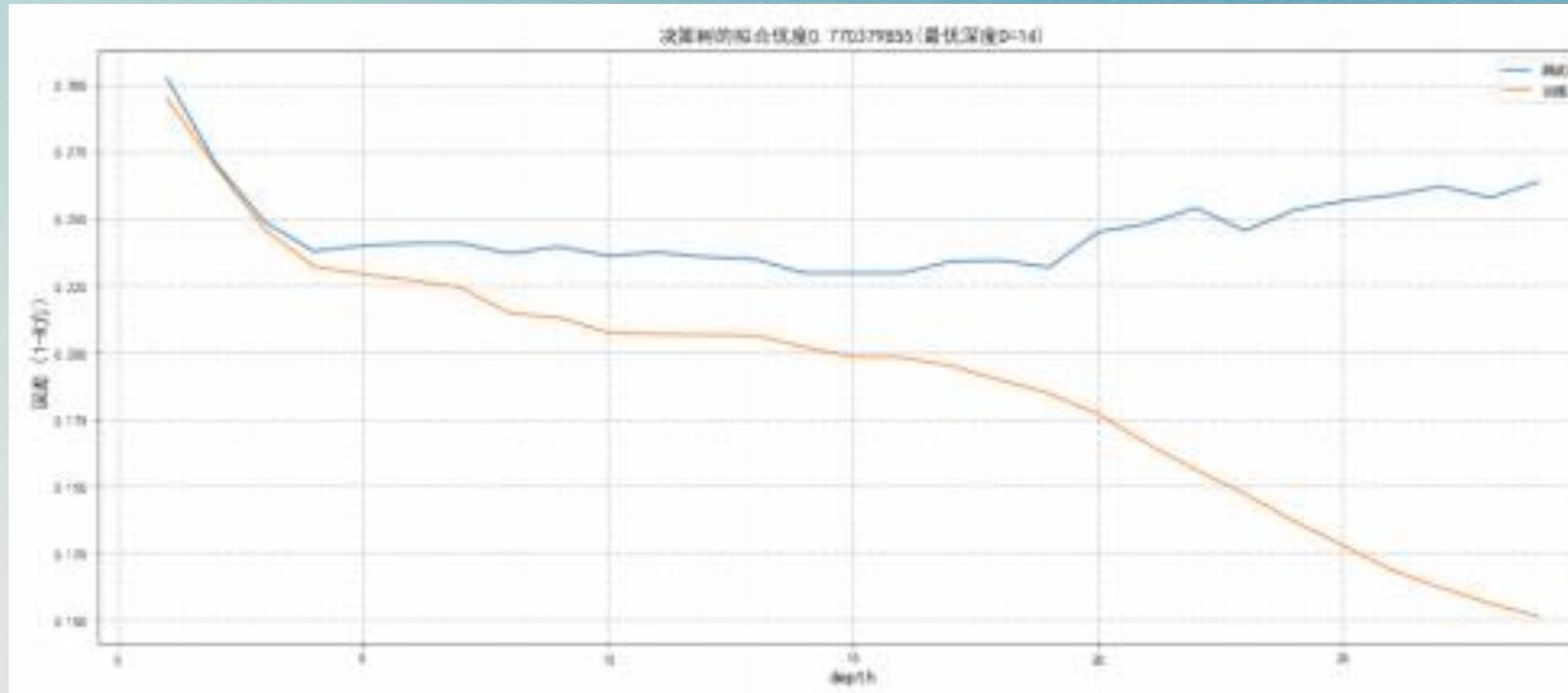
Adaboost集成学习

从上到下，从左到右依次为 I/E, N/S, F/T, P/J 的测试集评价结果，从结果可以看出，朴素贝叶斯的分类效果较为良好，其平均准确率达到了 0.647。



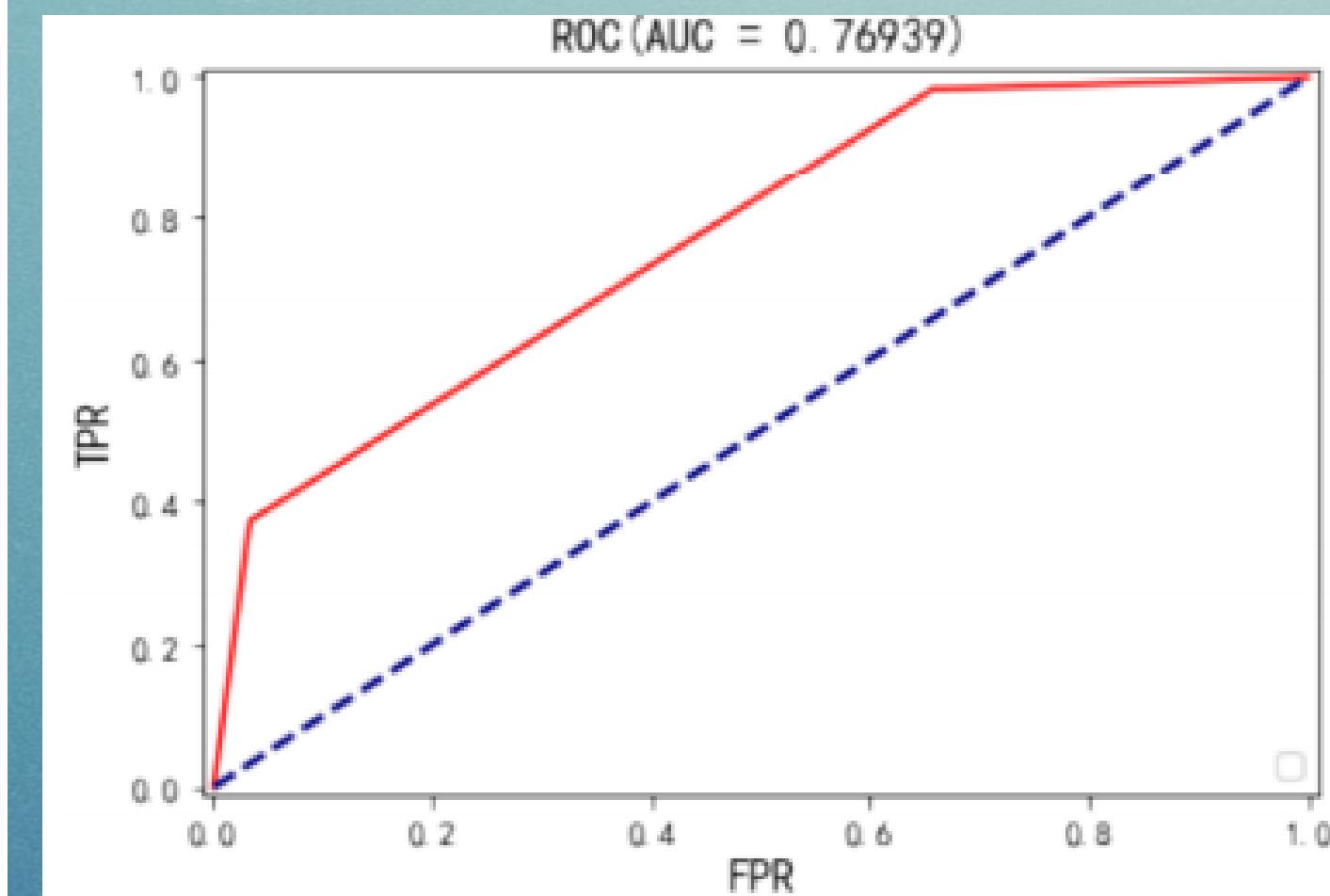
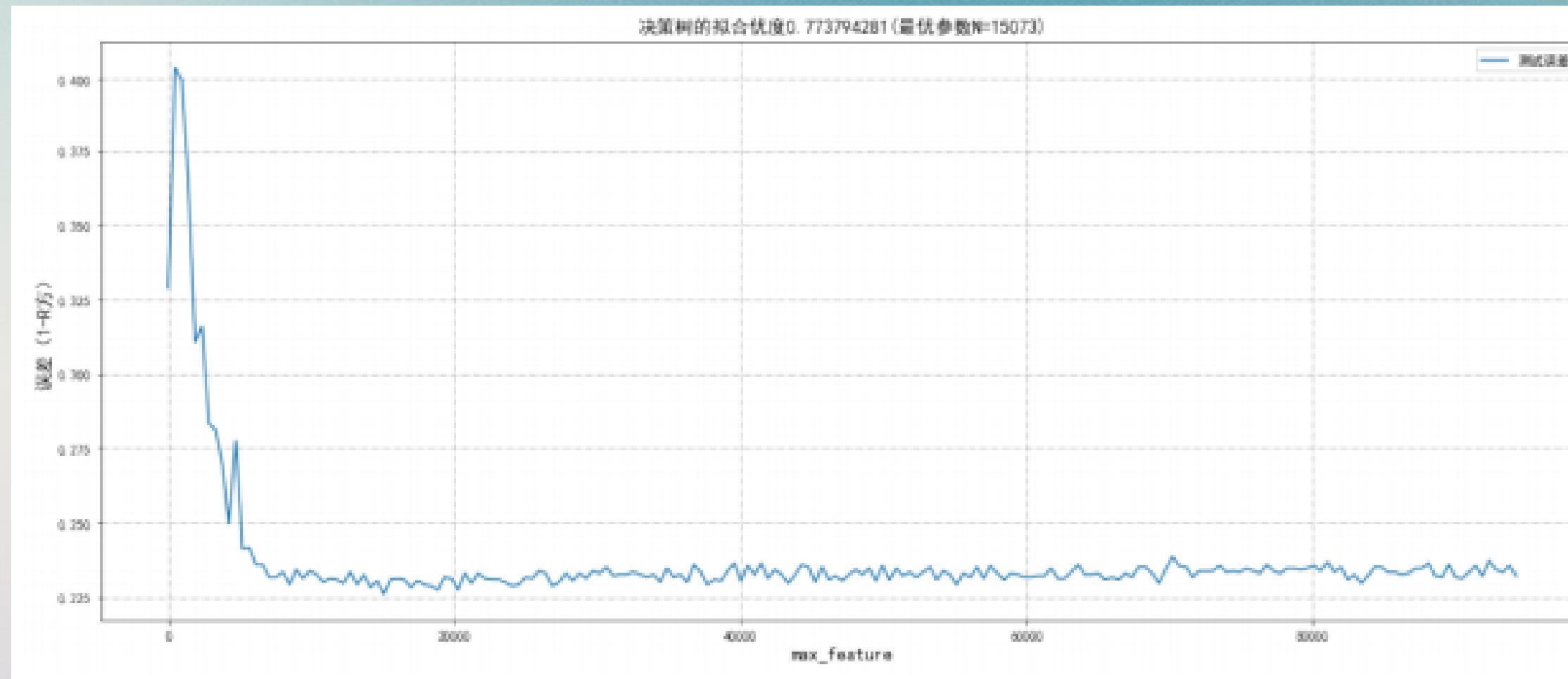
决策树

最优深度搜索：从上到下，从左到右依次为 I/E, N/S, F/I, P/J 的决策树最优深度



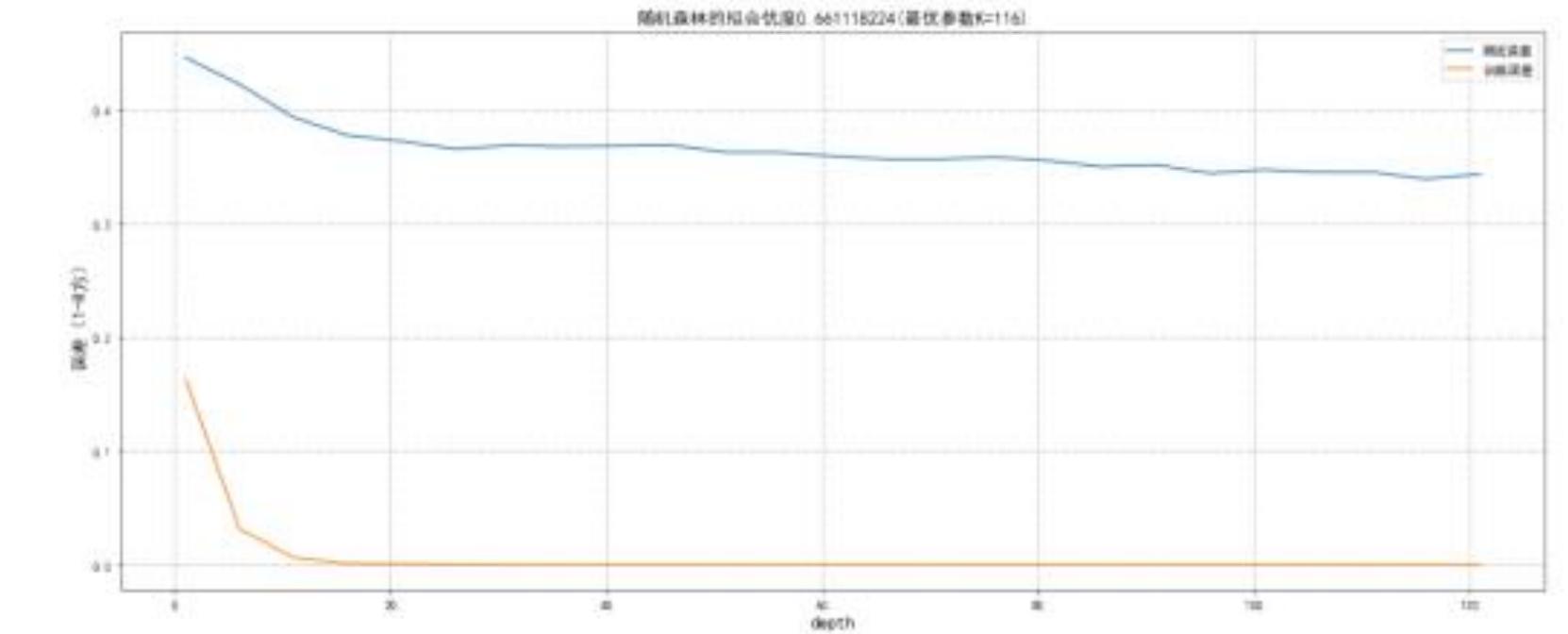
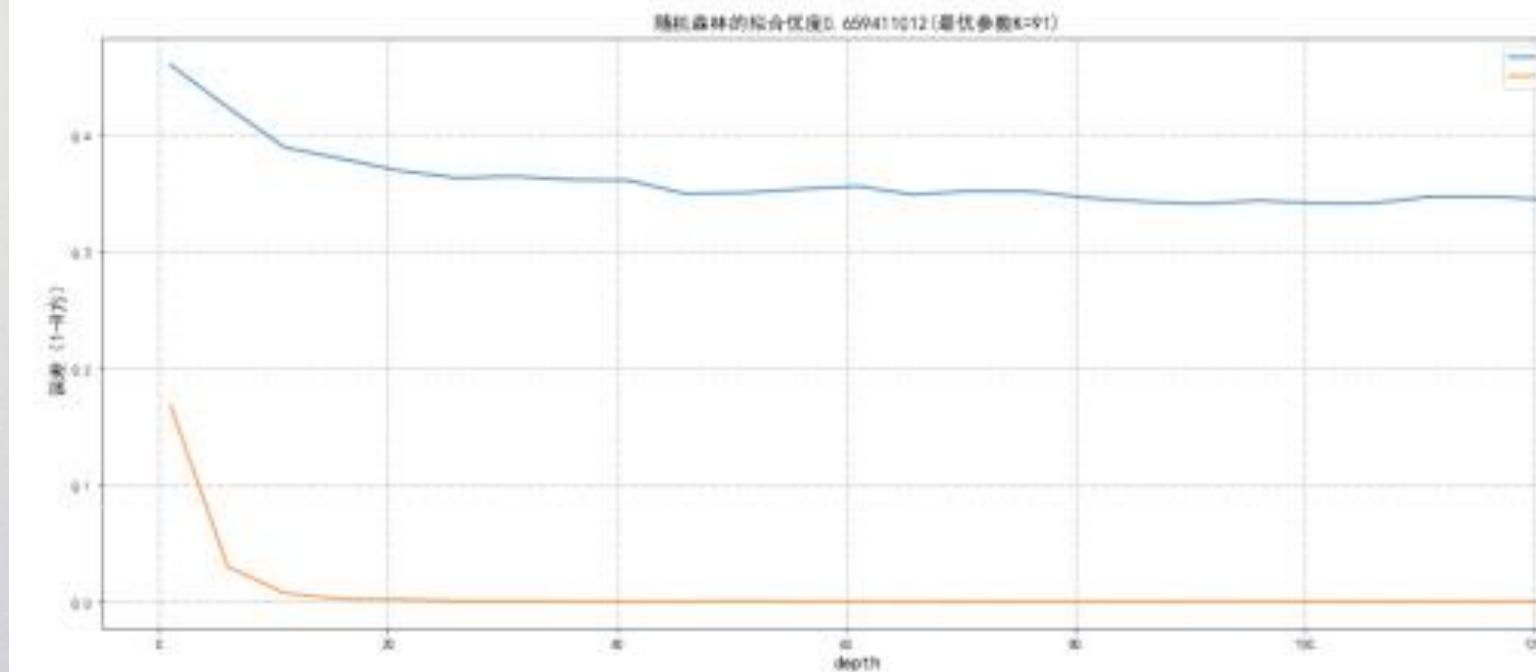
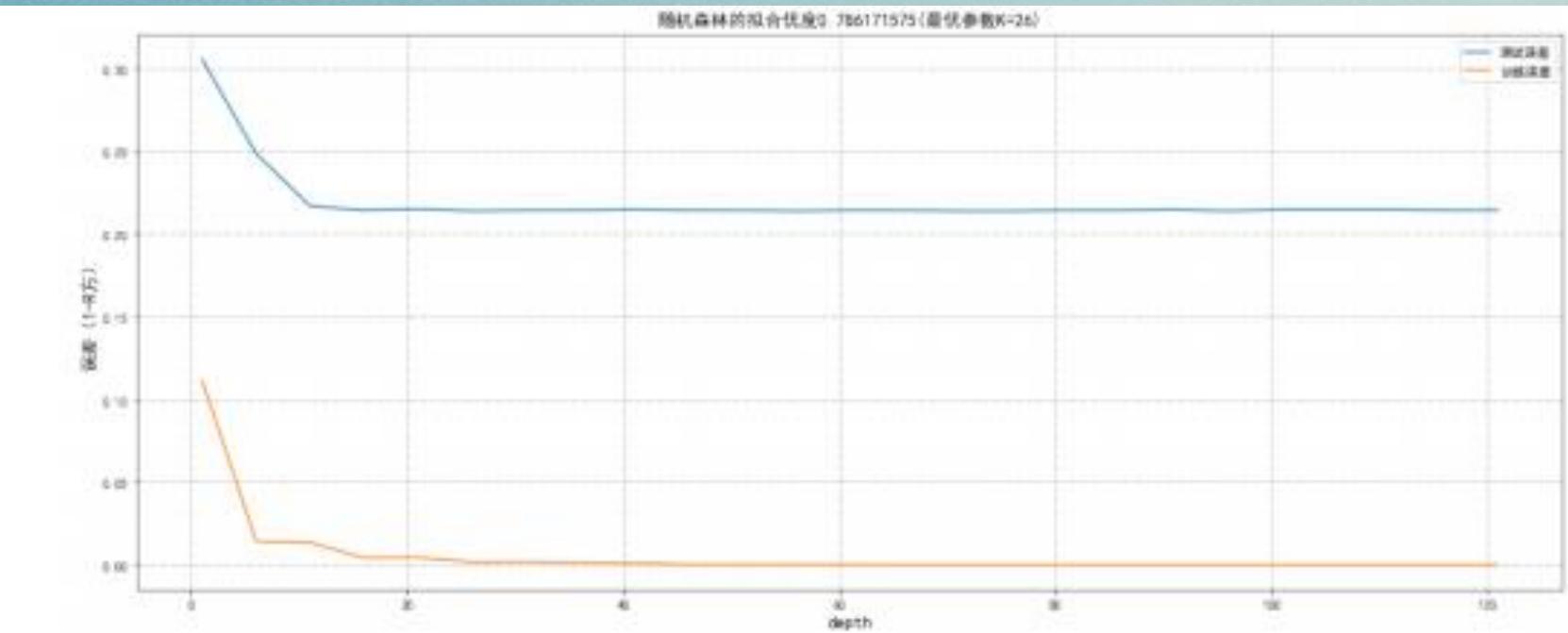
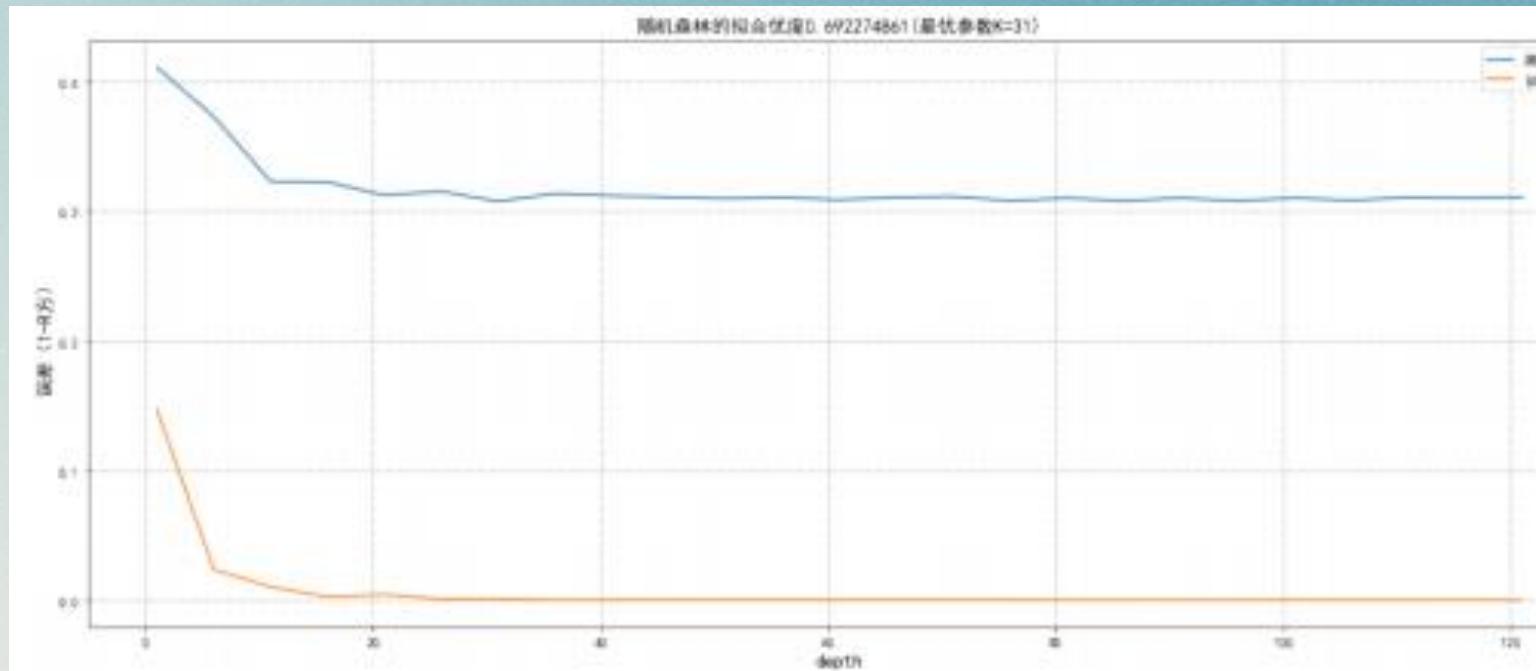
决策树

以I/E为例，在对Tf-Idf最优特征值数量和决策树最优深度这两个参数进行搜索后，输出的最优结果，以及其ROC曲线



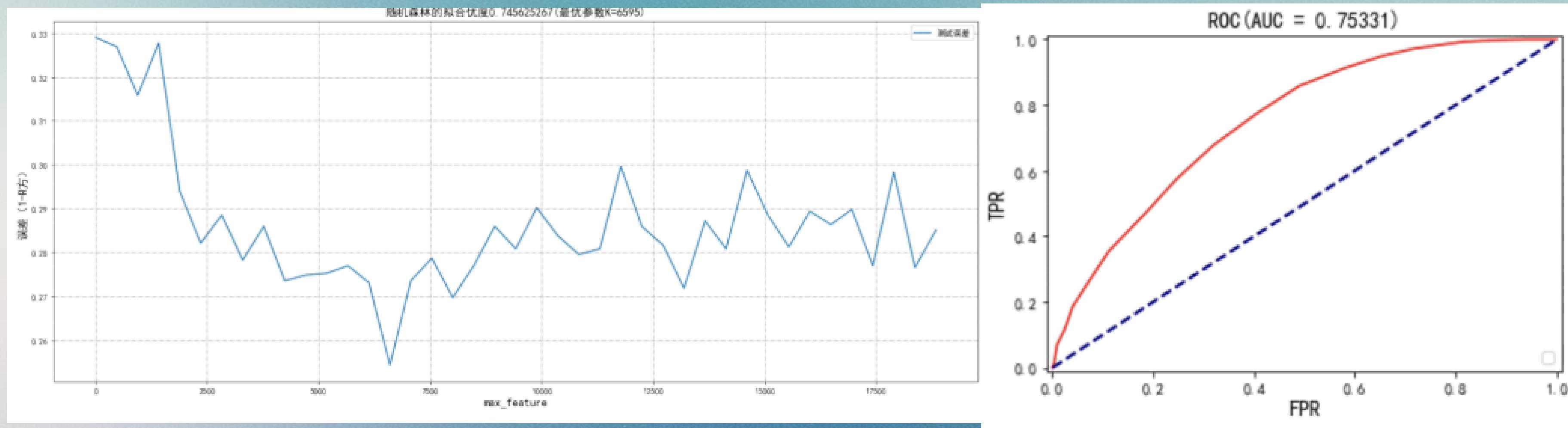
随机森林

最优深度搜索：从上到下，从左到右依次为 I/E, N/S, F/I, P/J的决策树最优深度

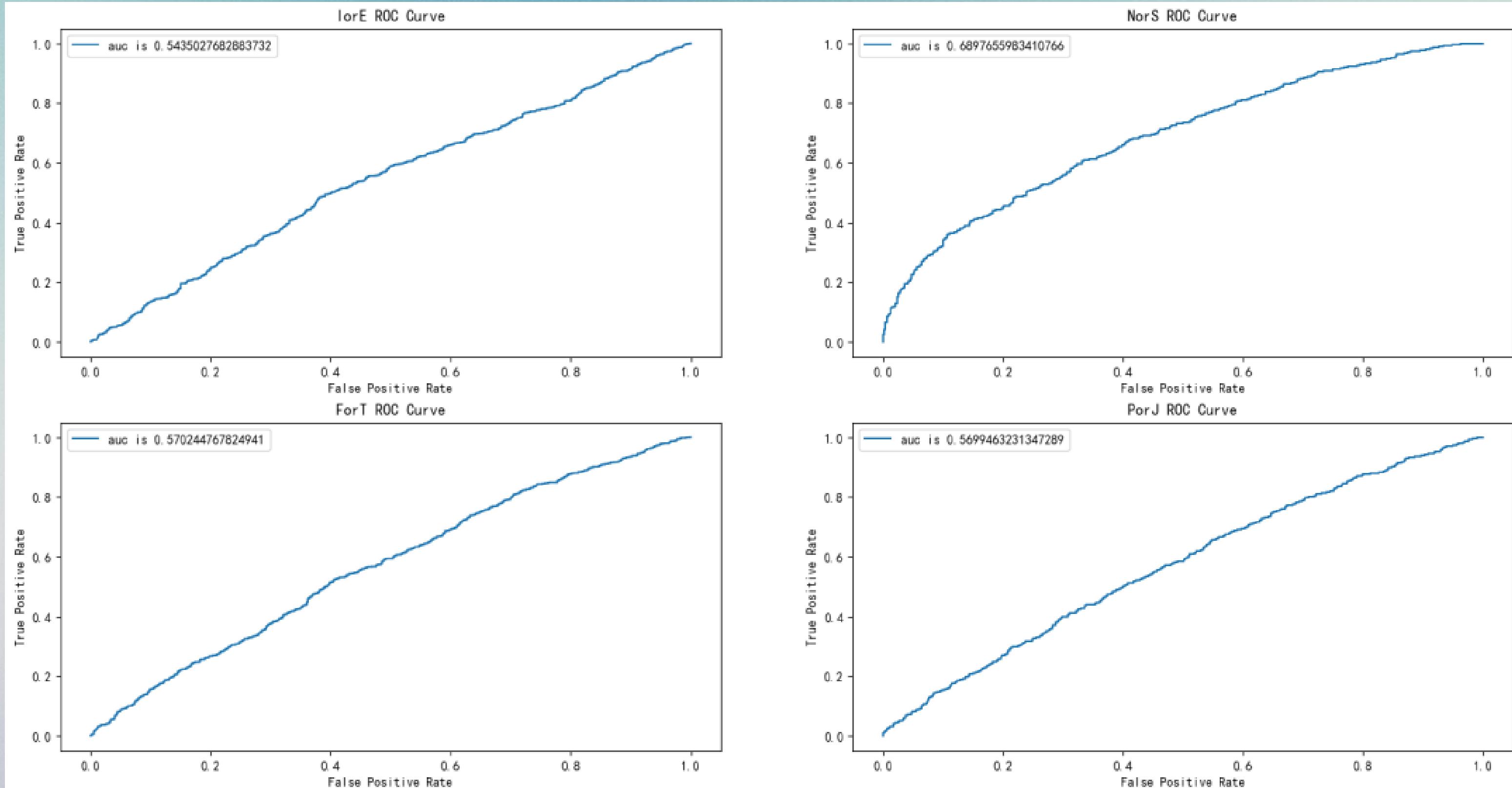


随机森林

以I/E为例，在对Tf-Idf最优特征值数量和决策树最优深度这两个参数进行搜索后，输出的最优结果，以及其ROC曲线

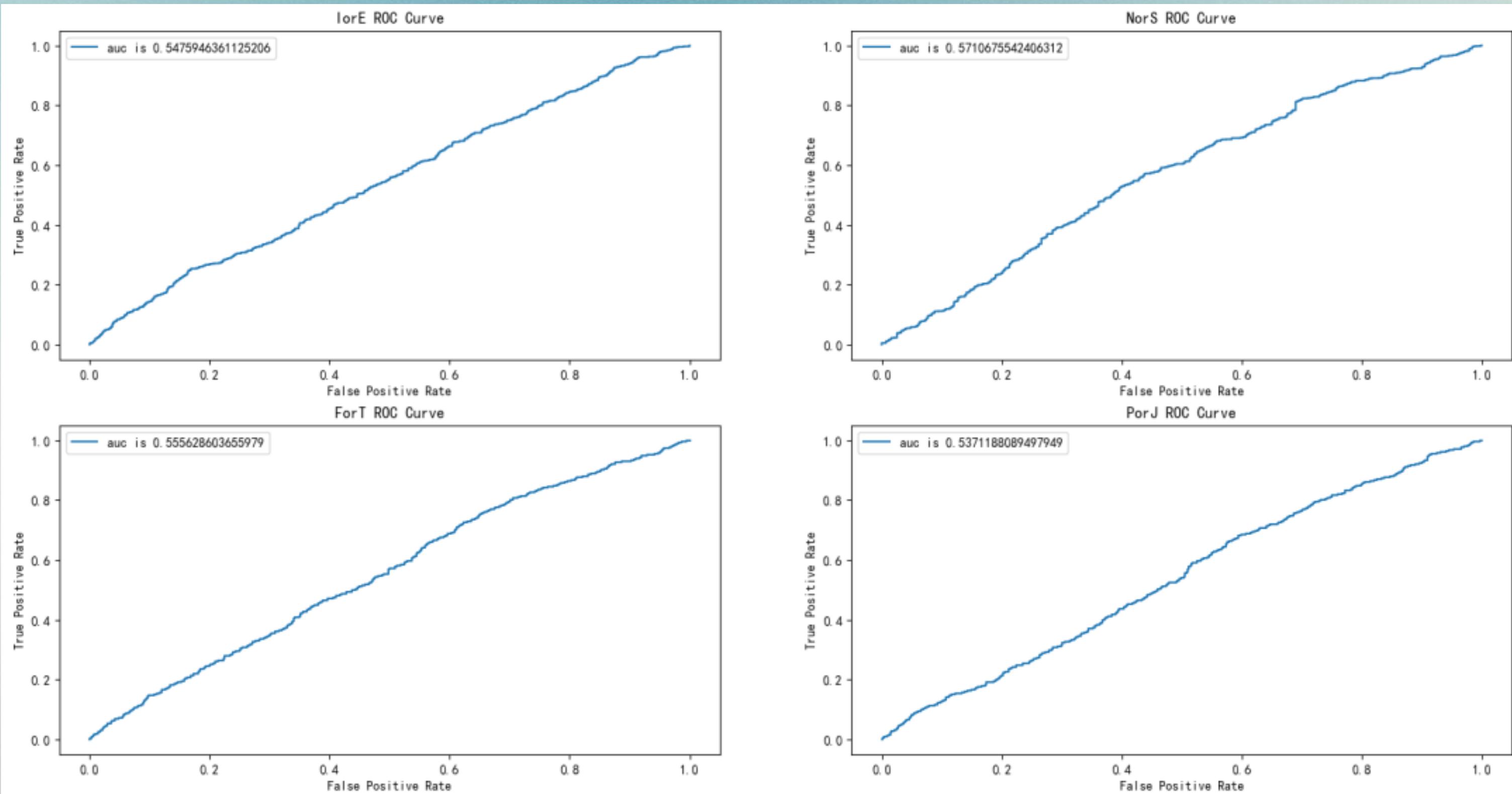


神经网络



单目标神经网络在四个特征维度上的分类效果

神经网络



多目标神经网络在四个特征维度上分别分类效果

IorE: Introversion (I) - Extroversion (E) ...
* IorE: Introversion (I) - Extroversion (E) Accuracy: 64. 08%
NorS: Intuition (N) - Sensing (S) ...
* NorS: Intuition (N) - Sensing (S) Accuracy: 70. 60%
ForT: Feeling (F) - Thinking (T) ...
* ForT: Feeling (F) - Thinking (T) Accuracy: 62. 34%
PorJ: Perceiving (P) - Judging (J) ...
* PorJ: Perceiving (P) - Judging (J) Accuracy: 64. 97%

线性支持向量机

简要说明您想要讨论的内容。

支持向量机

IorE: Introversion (I) - Extroversion (E) ...
* IorE: Introversion (I) - Extroversion (E) Accuracy: 68. 31%
NorS: Intuition (N) - Sensing (S) ...
* NorS: Intuition (N) - Sensing (S) Accuracy: 77. 19%
ForT: Feeling (F) - Thinking (T) ...
* ForT: Feeling (F) - Thinking (T) Accuracy: 61. 29%
PorJ: Perceiving (P) - Judging (J) ...
* PorJ: Perceiving (P) - Judging (J) Accuracy: 61. 09%

广义线性支持向量机

简要说明您想要讨论的内容。

写下您的主题或观点

| 写出列名称 | 写出列名称 | 写出列名称 | 写出列名称 | 写出列名称 |
|--------|-------|-------|-------|-------|
| 双击添加文本 | | | | |
| | | | | |
| | | | | |
| | | | | |
| | | | | |

讨论与总结

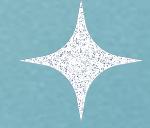
[返回目录页面](#)

写下您的 主题或观点

简要说明您想要讨论的内容。



感谢观看！



此处插入结束语或号召语。