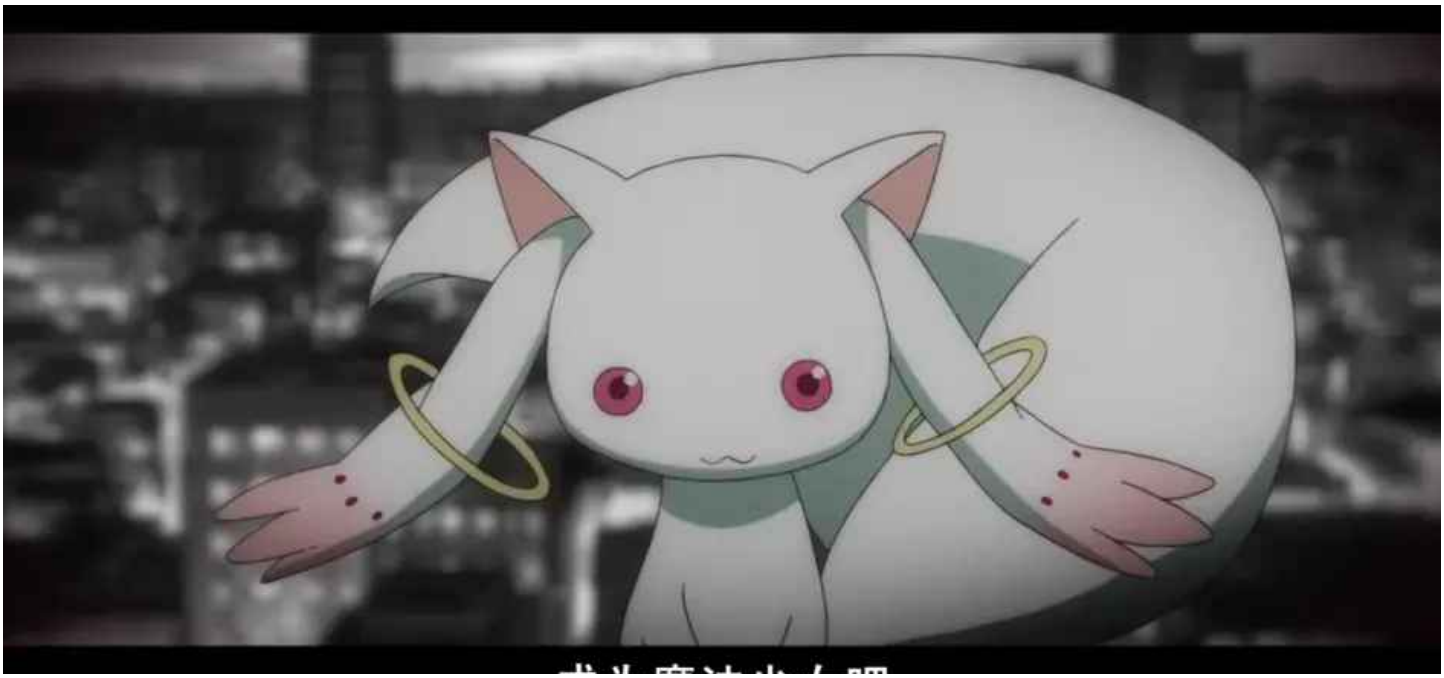


实验报告

I. 引言

介绍MBTI的概念和作用

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。它由凯瑟琳·布里格斯（Katharine Briggs）和伊莎贝尔·布里格斯·迈尔斯（Isabel Briggs Myers）母女所开发，并基于瑞士心理学家卡尔·荣格（Carl Jung）的人格理论。

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：

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

通过将这四对维度组合，MBTI定义了16种不同的人格类型，例如ISTJ、ENFP、INTP等。每种人格类型都有独特的特征和倾向，涵盖了个体的行为方式、沟通风格、决策方式和工作偏好等方面。

MBTI的作用在于提供一种框架来理解和描述个体的人格类型，帮助人们更好地认识自己和他人，同时在个人发展、职业规划、团队建设、人际关系管理等方面具有广泛的应用：

1. 职业发展：MBTI可以帮助个体了解自己的职业偏好和适合的工作环境，从而做出更明智的职业选择和规划。
2. 团队建设：通过了解团队成员的人格类型，可以促进团队合作、提高沟通效果和解决冲突，从而增强团队的工作效能。
3. 领导力发展：MBTI可以帮助领导者了解自己的领导风格和倾向，从而提高领导力的效果和适应不同的团队成员。
4. 人际关系管理：通过了解他人的人格类型，可以更好地理解他们的行为和需求，改善人际关系、提高合作和沟通效果。

如今MBTI被广泛使用，并被认为是一种有助于人们了解自己和他人的工具。通过评论数据辅助分析用户的MBTI类型与传统测评方法相比，传统的MBTI测试通常依赖用户的自我报告，而这种方法存在主观性和记忆偏差的问题。通过评论数据分析，可以避免这些问题，因为用户在评论中表达了真实的观点和反应，而无需依赖他们对自己的主观描述，为研究和实际应用提供更深入的洞察和理解。

介绍评论文本挖掘的背景和意义

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。这些方法结合机器学习和统计模型。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。

引入研究目的和研究问题

在本报告中，我们获取到不同人格类型的发言文本。通过应用文本挖掘方法，

II. 相关工作

回顾过去的研究关于MBTI和语言/评论文本之间的关系

讨论使用机器学习方法挖掘评论文本与MBTI之间关系的先前研究

预测模型分类评价指标

鉴于MBTI的特征，我们假设一个人格中的四对维度（性格取向，感知方式，决策方式，生活方式）相互独立，故我们对四对维度分别进行二分类预测，采用准确率（accuracy）作为二分类模型评价的指标，进一步综合四个维度的预测准确率，计算出MBTI 16人格的正确维度预测准确率，最终以此作为评价分类预测模型的指标。

III. 数据收集和预处理

数据来源

本研究所使用的数据集是基于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8711。

数据集的特征主要包括以下几个方面：

1. 用户推文数据：数据集中记录了每个用户的推文文本数据，这些推文是用户在某一段时间内，在社交媒体平台上发布的言论和观点。推文数据可以包含各种主题、情感和表达方式，反映了用户的思想、兴趣和交流方式。
2. MBTI人格类型：数据集中还包含了用户自报的MBTI人格类型信息。MBTI（Myers-Briggs Type Indicator）是一种常用的人格分类系统，通过将人格特质组合成16个类型，用以描述个体的行为偏好和认知方式。在数据集中，每个用户都提供了自己的MBTI人格类型，这可以用于与其推文内容进行关联和分析。

通过对这些数据进行深入分析，可以揭示不同人格类型用户在推文内容上的差异和共性，进一步探索人格与语言使用之间的关系。

描述数据预处理过程，包括文本清洗、标记化和特征提取等

在数据预处理阶段，我们使用了一系列文本处理技术来清洗和规范评论数据。以下是我们采取的主要步骤：

- 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。
- 文本转换为小写：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。
- 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。

在数据预处理的过程中，我们还提取了一些与情感分析相关的特征，包括：

- 表情数量：统计了每个评论中出现的表情数量，以衡量情感表达的强度和多样性。

- 文本长度：计算了经过处理后的评论文本的长度，以获取评论的文本信息量。
- 特殊字符数量：统计了每个评论中出现的特殊字符（如感叹号和问号）的数量，以衡量情感强度和评论的情感倾向。

同时考虑到用户文本的情感倾向有可能会与用户的人格特征有相关性，因此我们使用了NLTK库中的`SentimentIntensityAnalyzer`情感分析器，获取评论的情感评分。该分析器基于情感词典和规则，计算出评论的情感评分，包括积极情感、消极情感和中性情感。我们对每个评论进行了分句处理，并计算了每个句子的情感评分，然后取平均值作为整个评论的情感评分。

通过对数据进行预处理和文本清洗，我们可以减少噪声、规范文本格式，提取有用的特征，为进一步实现文本编码功能做好铺垫。其中表情转换、文本转换为小写、去除网址和非英文字符、词性还原和分词等处理步骤可以帮助我们获得更干净、一致的文本数据。提取的特征如表情数量、文本长度和特殊字符数量可以提供更多的信息来解释和分析评论的情感倾向。然而，我们也注意到数据预处理和文本清洗可能会引入一些偏差或信息损失。例如，在去除停用词的过程中，可能会丢失一些具有情感倾向的停用词，因此需要在具体任务中进行权衡和调整。

文本词编码过程

词向量Word2vec编码

Word2vec的工作原理基于神经网络语言模型，它将词语表示为向量，并学习语料库中的词汇之间的相关性和语义关系。Word2vec通过分析语料库中的词语序列来学习这些关系，并使用这些关系来预测语料库中的下一个词语，这样很好的保留了单词词义的信息。在训练过程中，Word2vec会更新每个词语的向量来更好地表示它们的语义。最终，Word2vec会生成一个词语向量空间，其中每个词语都被表示为一个向量，这些向量可以用于词语相似度计算，语义分类等语言处理过程。

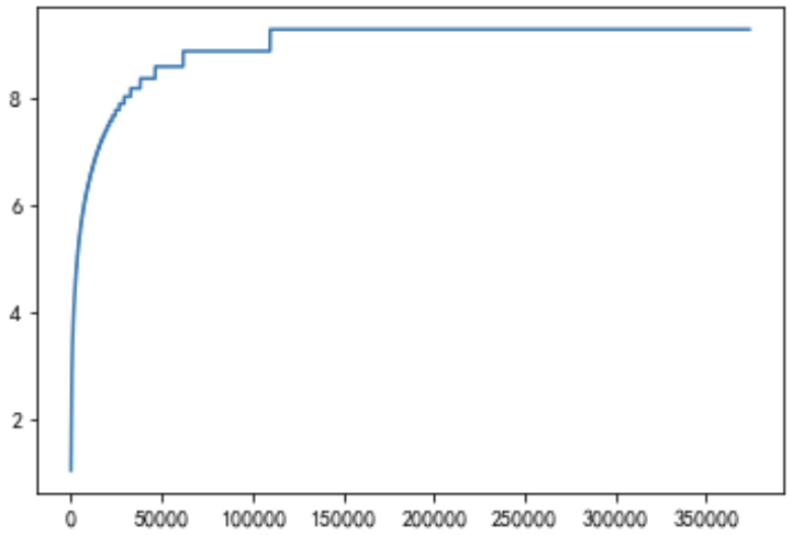
在Word2vec模型中，根据词的上下文我们采用skip-gram算法进行编码训练，我们采用数据集中7000多条评论作为训练数据，采用5作为一个句子中当前单词和被预测单词的最大距离，最终输出每个词对应一个100维的向量编码，对于数据集中7811条评论数据，经过相关资料查阅，我们认为采用词向量中的最大值作为该词语的编码输出是合适的，同时采用每条评论中的前400个词语作为句子特征进行句子编码，如果评论词数少于400词则用0填充，最终得到7811x400维的训练集向量。

我们用词语间的相似度合理性作为评价Word2vec模型的指标，该模型通过计算两个词语间的词向量相似度来计算词语间的相似度，比如我们输出与词语'bread'最相似的五个词
为'cheese'，'soup'，'sandwich'，'chip'，'chocolate'，他们都共同有着食物这一语义含义，这很好的反映了Word2vec模型保留词义特征的特点。

Tf-Idf编码

TF-Idf的工作原理基于词频和逆文档频率的乘积，它将词语表示为权重，并衡量词语对于文档的重要性。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，这样很好地提取了文档的关键词。在计算过程中，TF-Idf会调整每个词语的权重来更好地反映它们的重要程度。最终，TF-Idf会生成一个词语权重空间，其中每个词语都被表示为一个权重，这些权重可以用于文档相似度计算，信息检索等文本挖掘过程。

在Tf-Idf模型中，为确保不加偏袒地反映出每个人格对应文本包含的信息，我们首先对整个文本进行了统一的Tf-Idf编码，得到了7810*373891维的训练向量，该训练向量将用于后续的研究。下图为Tf-Idf在关于idf排序后绘制的分布图。横坐标代表累积向量个数，纵坐标代表向量的Tf-Idf值。



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，
如：'able'，'crazy'，'fear'，'worry'等等，这侧面验证了Tf-Idf模型的可靠性。在提取相似性矩阵之后，我们发现一些相似性在99%以上的文本，这对一些人格“喜欢发大量重复性的语句”这一特征进行了很好地反映，同样证明了Tf-Idf模型的合理性。

在后续的数据分析中，由于各个人格向量是相互独立的，为获得每个人格向量上的最优特征数数量，我们利用for循环进行搜索（由于涉及到两个模型，故无法使用Grid SearchCV），并将最优特征数数量用于各自人格向量的结果输出。

词袋BOW编码

BOW的工作原理基于词频，它是将文档表示成特征矢量。它的基本思想是假定对于一个文本，忽略其词序和语法、句法，仅仅将其看做是一些词汇的集合，而文本中的每个词汇都是独立的。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，将两篇文本通过词袋模型变为向量模型，通过计算向量的余弦距离来计算两个文本间的相似度。

在词袋模型中，由于总体语料库往往非常庞大，很容易导致生成的向量维度很高且稀疏，所以在实际处理的过程中，还会进行词频统计这一步，即统计每个词在数据集中出现的次数，然后只选择其中出现频率最高的前5000个词作为最终的词表。这能反映出某种类型的人对哪些词的使用频率更高，也符合MBTI测试的概念。

IV. 方法和实验设计

解释所使用的机器学习方法，如情感分析、主题建模等

详细描述实验设计，包括训练集、验证集和测试集的划分，特征选择等

讨论所用评估指标，如准确率、召回率、F1值等

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。

我们这里采用词向量进行编码后，将数据集按照7:3的比例划分为训练集和测试集，采用sklearn中的MultinomialNB模型进行贝叶斯二分类预测任务。

KNN

选择KNN的原因

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。

我们这里采用词向量进行编码后，在1-50的K值范围内搜索最佳参数K，使得预测误差在训练集上达到最小。

Adaboost集成学习

选择Adaboost集成学习的原因

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。

我们这里用Word2vec模型对评论编码后采用简单决策树模型作为弱学习器来建立Adaboost集成学习框架，来完成人格维度的二分类任务。

决策树

选择决策树的原因

1. TF-IDF编码将生成大量的特征向量（高达7810*373891维）词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。
3. 决策树的结构可以直观地展示出特征之间的关系和判断过程，可以生成可解释性强的规则，便于理解和解释。
4. 决策树模型可以处理分类变量和连续变量，不需要做太多的数据预处理，例如不需要对数据进行标准化或归一化，也不需要对缺失值进行填充或删除。
5. 决策树通过递归地划分数据集，可以捕捉到数据中的非线性关系和交互效应，适用于复杂的数据分布，对异常值和噪声不太敏感。

逻辑斯特回归

逻辑斯特回归是一种线性判别式模型，可以通过对数几率函数将文本特征与0-1变量之间的关系建立起来，输出每个类别的概率估计。

选择逻辑斯特回归的原因

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。

随机森林

随机森林模型是由多个决策树组成的集成学习方法，可以充分利用决策树的优点，同时通过投票或平均的方式，降低单个决策树的方差和偏差。

选择随机森林的原因

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。
4. 随机森林模型可以处理高维度和大规模的文本数据，不需要进行特征选择或降维。

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。

选择TextCNN原因

1. TextCNN利用一维卷积神经网络可以有效地捕捉文本中的局部特征。由于文本数据通常具有局部的语义结构，卷积操作可以在不同大小的窗口上提取出不同级别的特征，从而更好地捕捉文本的语义信息。这种多尺度的卷积核能够捕捉不同长度的文本片段，并通过池化操作将它们合并在一起，形成更全局的特征表示。这样可以提高模型对文本的整体理解能力。在进行卷积操作时TextCNN实现了参数共享，使得模型的参数量较小且共享了相似的特征提取权重，减少了模型的复杂度，并且在训练过程中能够更快地收敛和优化，节省了训练时间和计算资源。相比于其他的文本分类神经网络结构，TextCNN的结构相对简单，并且可以通过可视化卷积核的响应来理解模型学习到的特征。这使得我们能够更好地理解模型的工作原理，并解释模型对文本分类的决策依据。
2. 与TextRCNN相比，一般TextRCNN模型在模型初始层加入LSTM可以帮助模型联系前后文本信息，但是在本模型中，因为我们使用了BERT模型作为我们的嵌入层（embedding）综合处理文本的句义，因此LSTM层并不是必需的。同时使用LSTM会大大增加模型训练以及运行的时间以及内存的开销，但是在正确率上，与TextCNN模型相比并没有较大提升（约增长0.01）。

通过多个二分类模型实现多目标分类任务

我们定义了三个卷积层，分别使用不同大小的卷积核来捕捉评论矩阵从微观到宏观不同尺寸的特征。这三个卷积层的输入均为上一步得到的输入张量。对于每个卷积层，我们指定了卷积核的数量和大小，并采用ELU激活函数对卷积层的输出进行非线性变换。在每个卷积层之后，我们添加了对应的池化层，以降低特征维度和提取最重要的特征。这里我们采用了最大池化操作，分别对应于每个卷积层的输出。在拼接层，我们将三个池化层的输出拼接在一起，形成一个更丰富的特征表示。这里我们使用Keras提供的Concatenate函数，并指定拼接的轴为1。在展平层，将拼接层的输出展平为一维向量，以便后续的全连接层处理。我们使用Keras提供的Flatten函数来实现展平操作。

最后，我们添加一个全连接层作为输出层，它的激活函数为sigmoid，输出二分类的概率值。在训练过程中考虑到数据不平衡对模型分类可能造成一定的影响。因此，我们使用了Focal Loss函数作为损失函数。Focal Loss通过引入类别权重的概念，可以有效地处理类别不平衡问题，使得模型更加关注难以分类的样本。

多目标分类模型

我们定义了三个卷积层，分别使用不同大小的卷积核来捕捉不同尺寸的特征。这三个卷积层的输入均为上一步得到的输入张量。对于每个卷积层，我们指定了卷积核的数量和大小，并采用ELU激活函数对卷积层的输出进行非线性变换。将三个卷积层的输出拼接在一起，形成一个更丰富的特征表示。这里我们使用Keras提供的Concatenate函数，并指定拼接的轴为1。展平层将拼接层的输出展平为一维向量，以便后续的全连接层处理。我们使用Keras提供的Flatten函数来实现展平操作。输出层我们为每个目标定义了一个独立的输出层。在这里，我们使用Dense层，并给出相应的单元数量。每个输出层都可

以输出一个连续值，表示相应目标的分类结果。通过将输入和多个输出层组合在一起，我们定义了整个模型的结构。最后，我们使用Keras提供的Model函数定义了完整的模型，给出输入和输出。

支持向量分类机

选择支持向量分类机的原因

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。

我们这里采用词袋模型进行编码后，将数据集按照2:1的比例划分为训练集和测试集，采用sklearn中的svm模型，调用不同的参数进行支持向量机分类。

V. 实验结果和分析

展示实验结果，包括模型性能指标和可视化分析

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

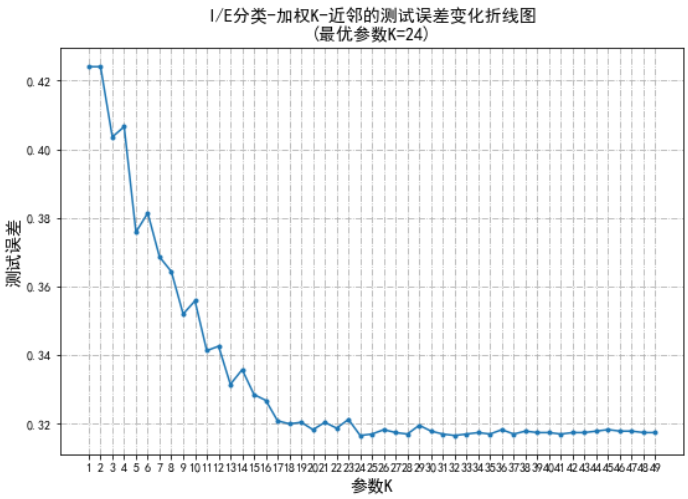
朴素贝叶斯分类器的测试误差:0.439
测试样本的评价结果:

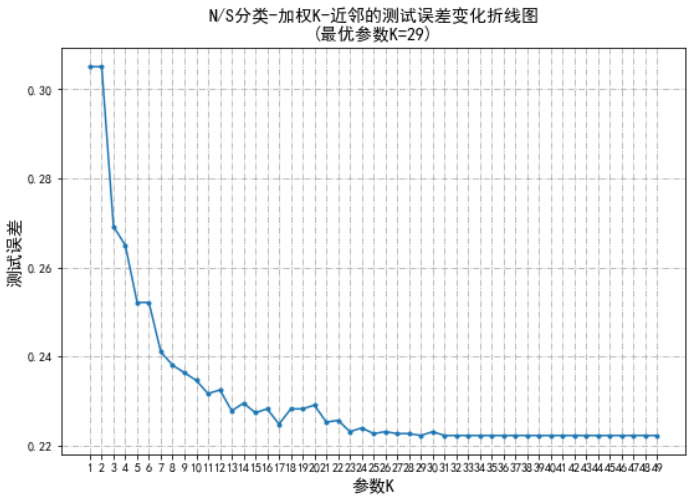
	precision	recall	f1-score	support
j	0.00	0.00	0.00	1027
p	0.56	1.00	0.72	1317
accuracy			0.56	2344
macro avg	0.28	0.50	0.36	2344
weighted avg	0.32	0.56	0.40	2344

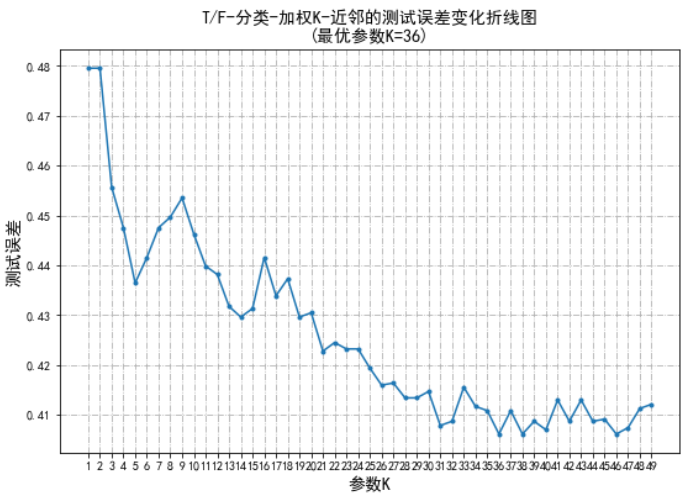
分析原因可以发现，朴素贝叶斯分类器判断出了几乎所有的多数类样本，但放弃了全部的少数类样本，很容易受到样本不均衡问题影响，又因为该数据文本向量化后特征值太多且相互之间存在相关性，将特征值二值化的选择并不优秀，故在此问题上该模型表现较差。

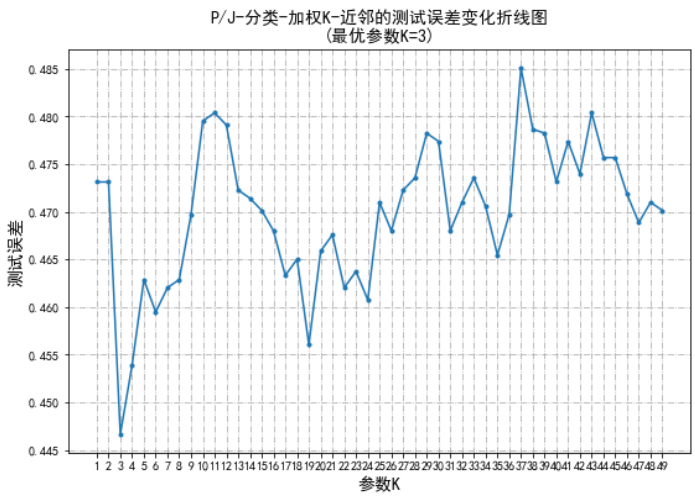
KNN

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



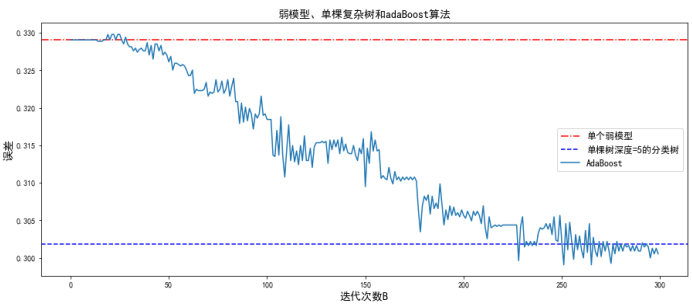


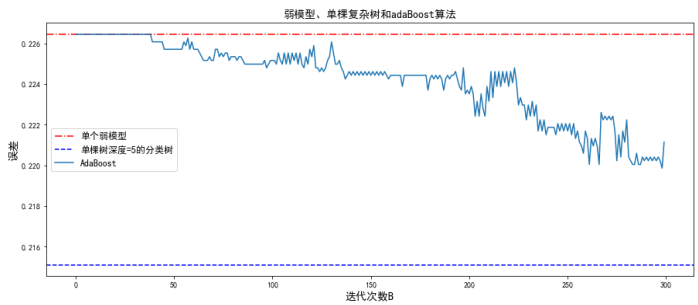


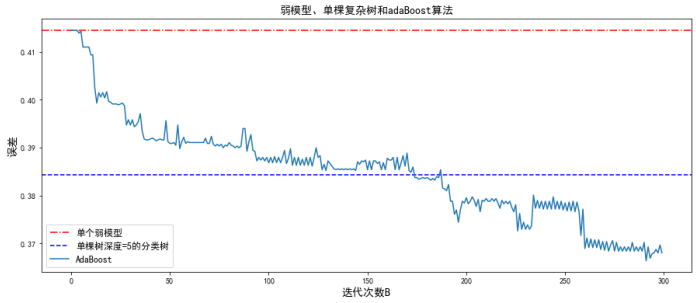


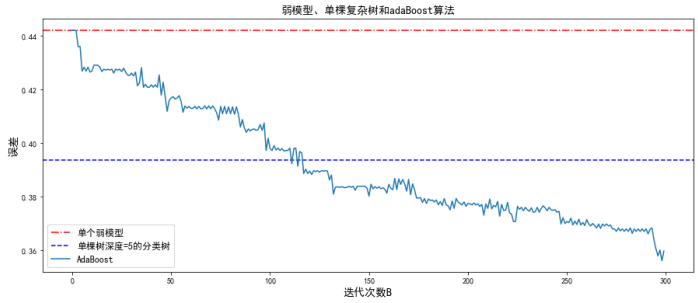
Adaboost集成学习

从上到下，从左到右依次为 I/E, N/S, F/T, P/J的训练误差的变化情况，综合看出深度为1的决策树弱模型误差较大，但随着迭代次数的增大，采用boost方法对弱模型的逐渐更新与建立，最终得到误差更小的强模型，综合来看，Adaboost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0.6325。

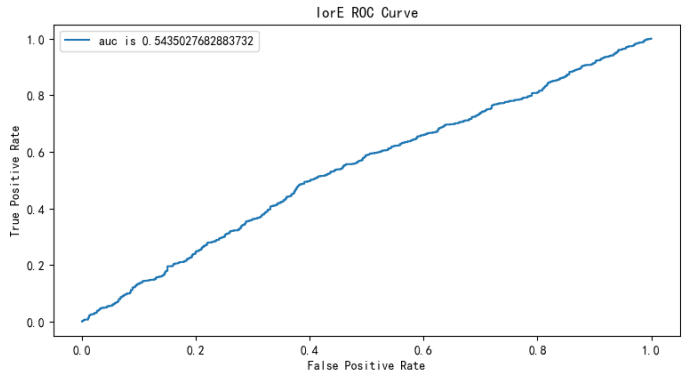


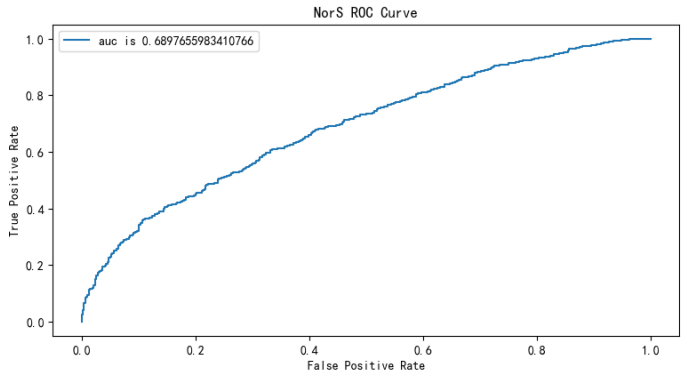


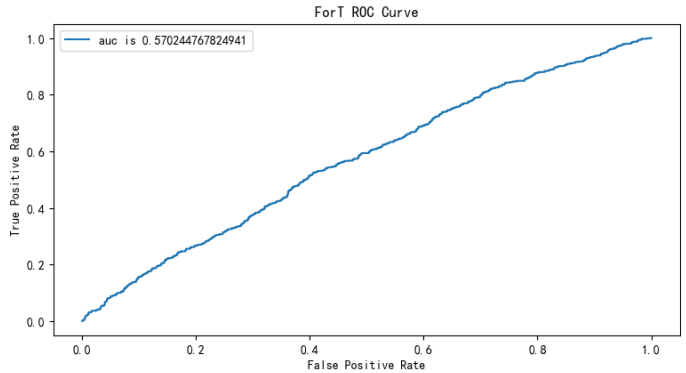


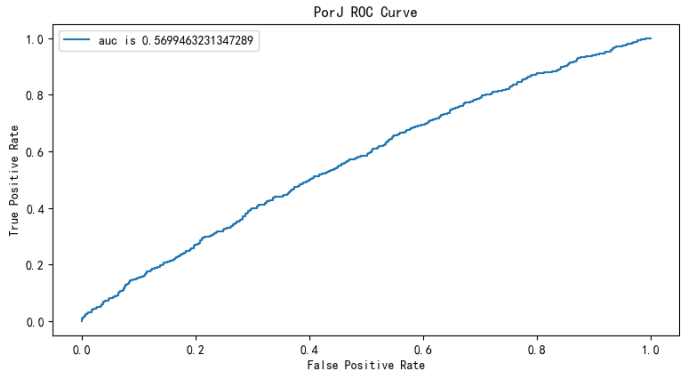


神经网络

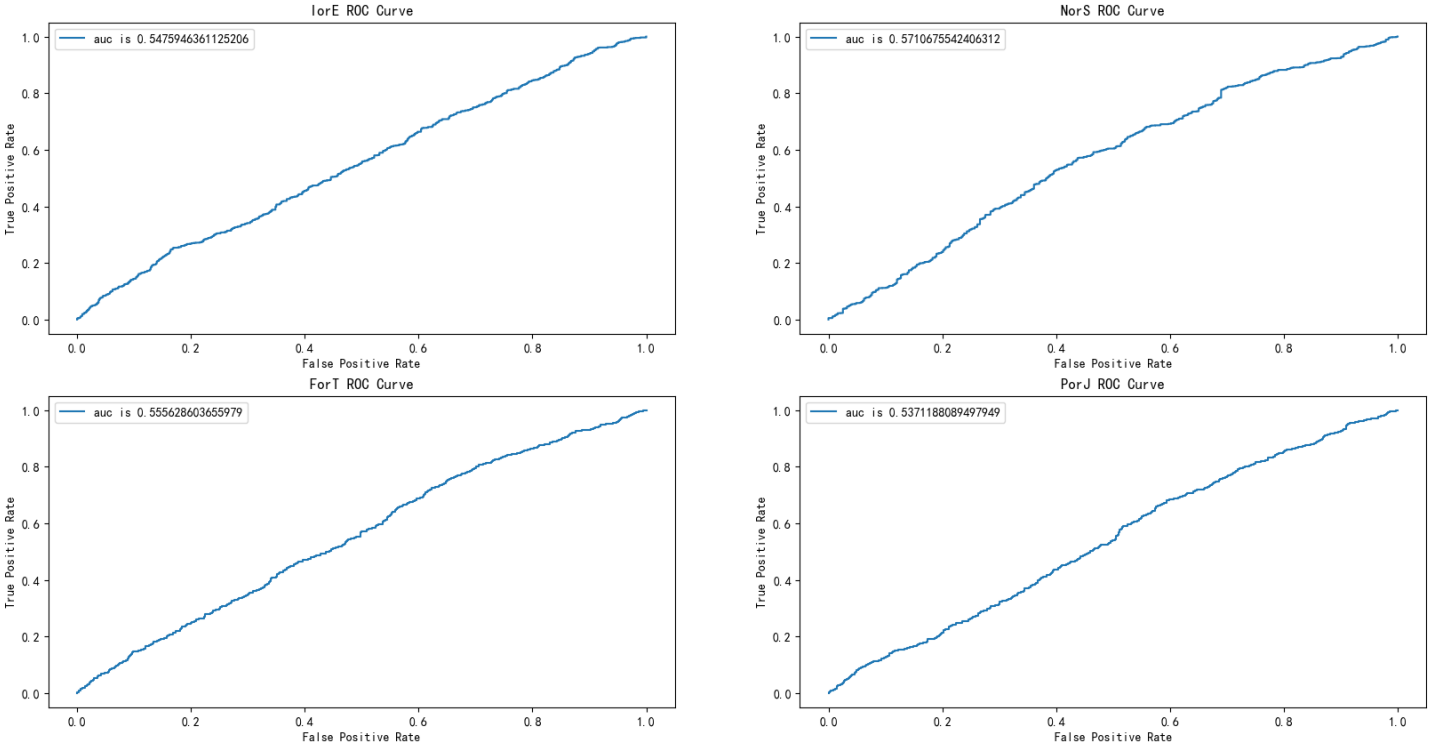








单目标神经网络在四个特征维度上的分类效果

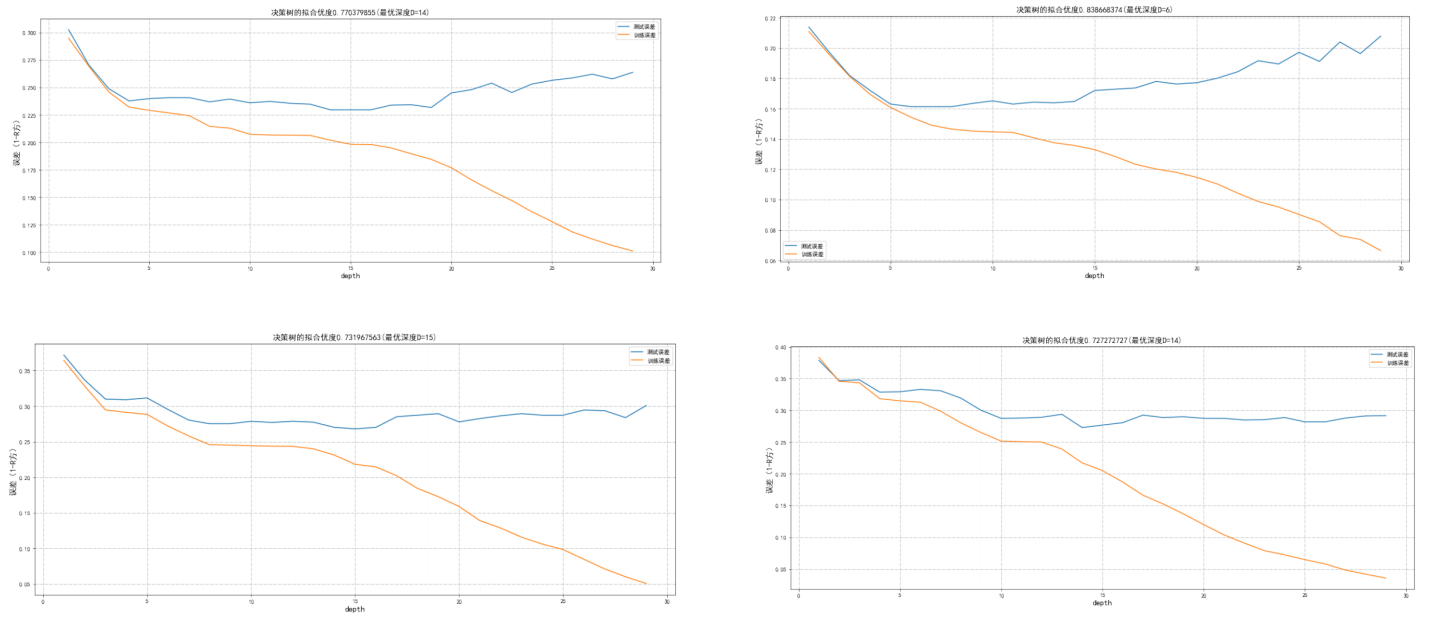


多目标神经网络在四个特征维度上分别分类效果

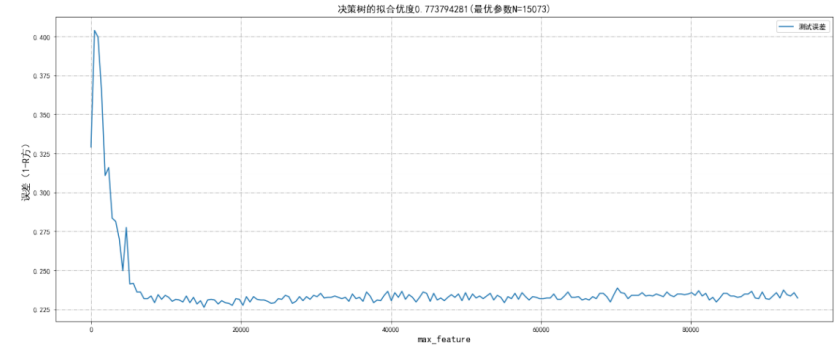
决策树

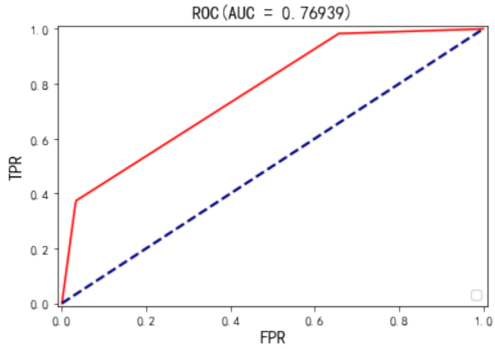
最优深度搜索过程：

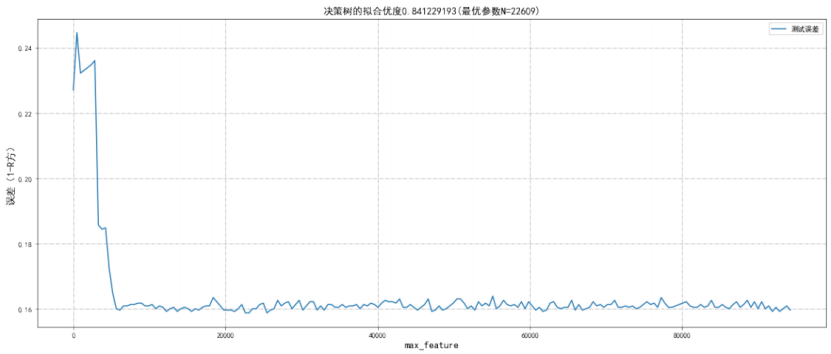
从上到下，从左到右依次为 I/E, N/S, F/I, P/J的决策树最优深度，这些在后续的分析里都会直接用于模型之中

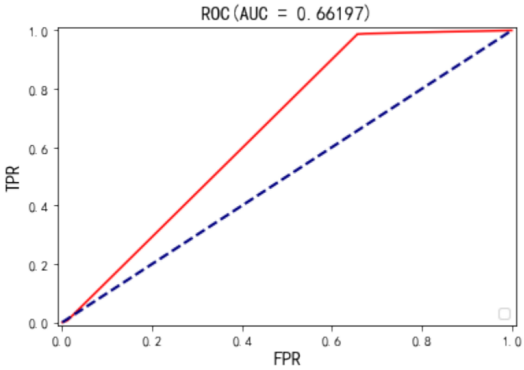


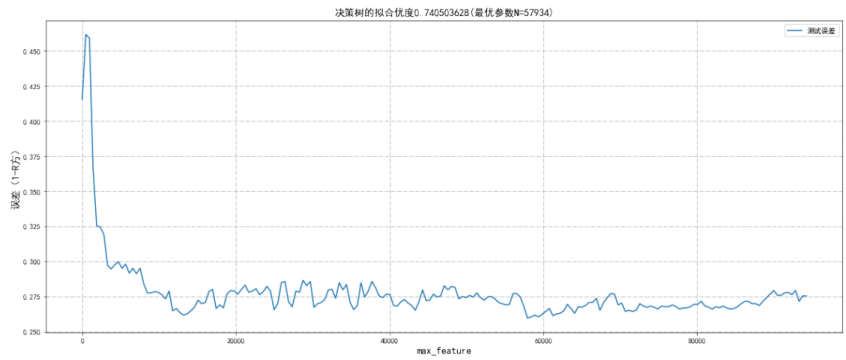
从上到下依次为I/E, N/S, F/T, P/J在对Tf-Idf最优特征值数量和决策树最优深度这两个参数进行搜索后，输出的最优结果，以及其ROC曲线

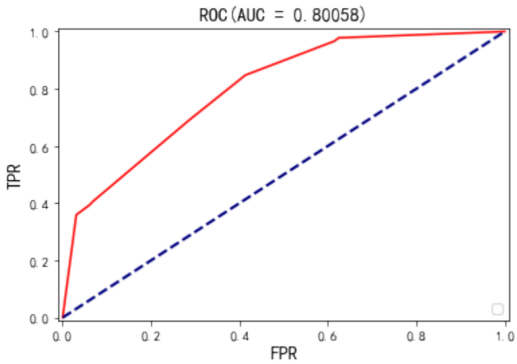


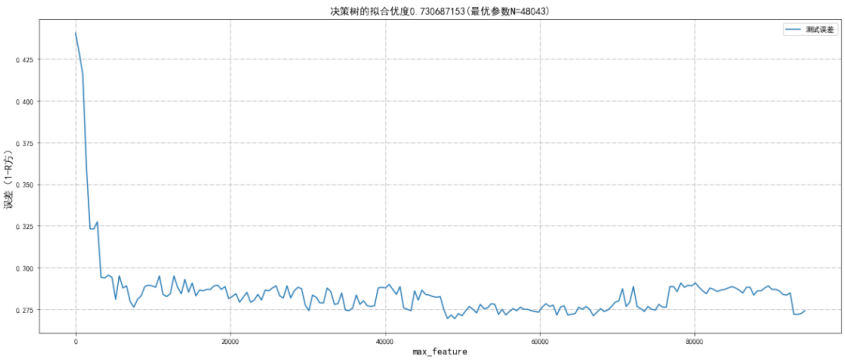


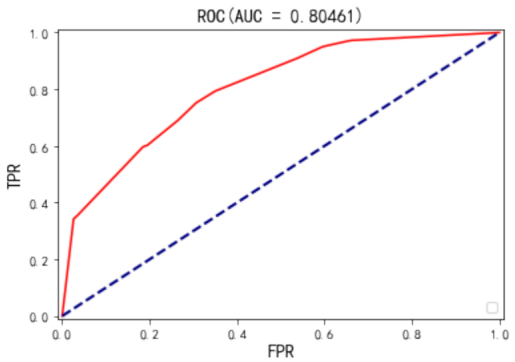










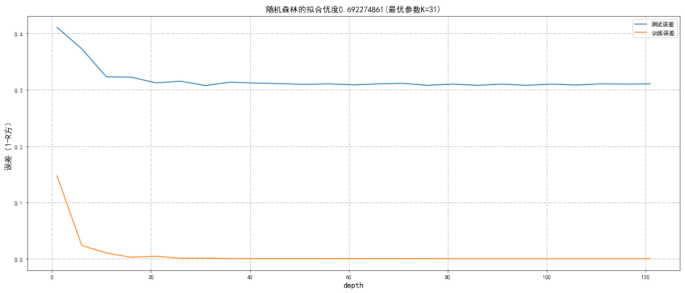


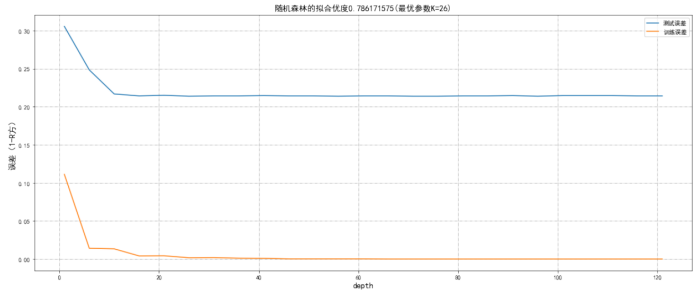
0.7718736662398634

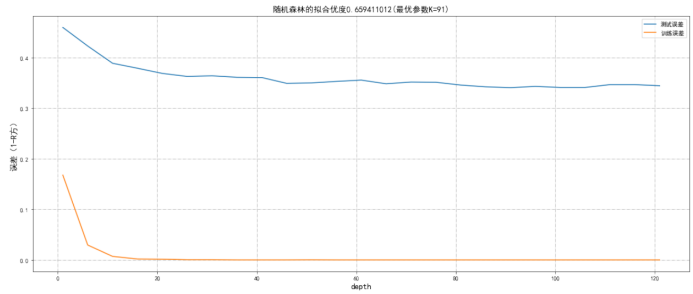
随机森林

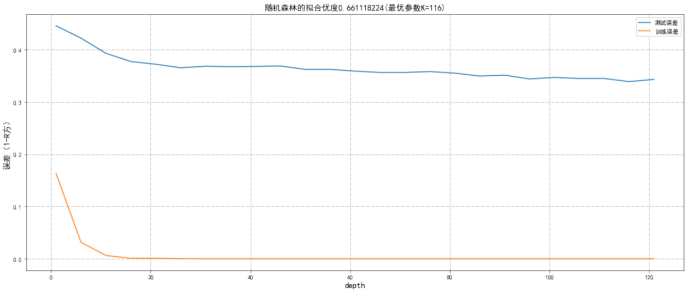
最优深度搜索过程：

从上到下，从左到右依次为 I/E, N/S, F/T, P/J的决策树最优深度，这些在后续的分析里都会直接用于模型之中

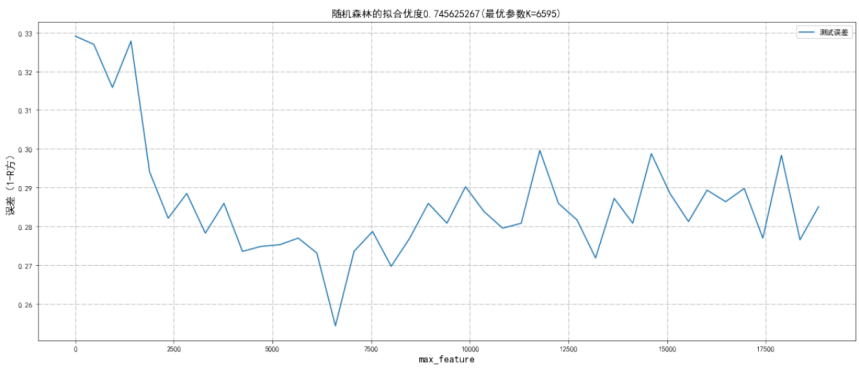


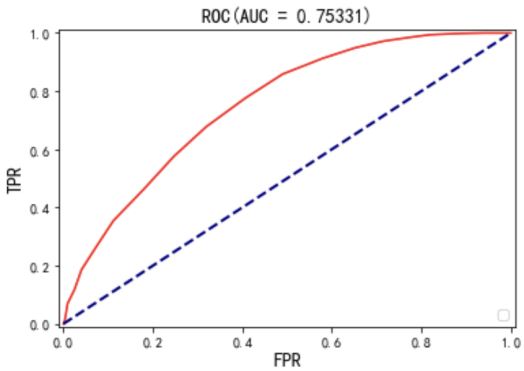


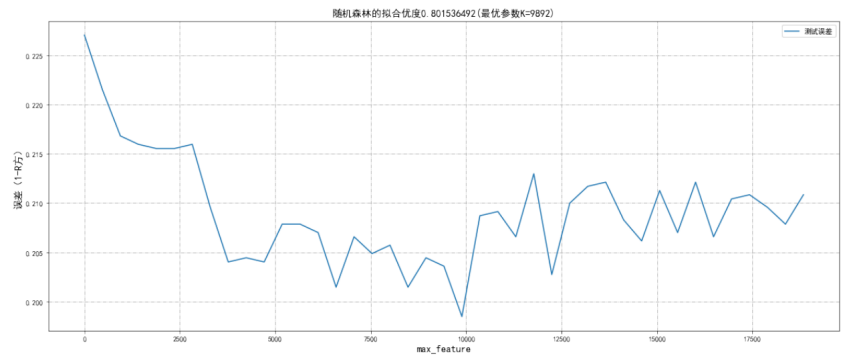


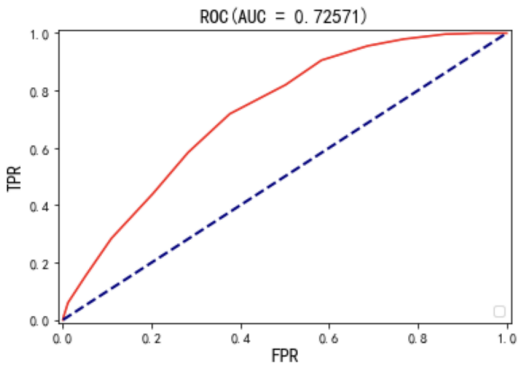


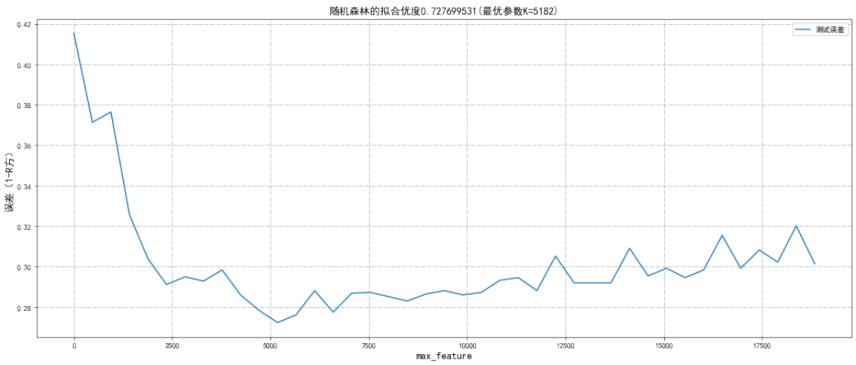
从上到下依次为I/E, N/S, F/I, P/J在对Tf-Idf最优特征值数量和随机森林最优深度这两个参数进行搜索后，输出的最优结果，以及其ROC曲线。可以看出，相较于决策数模型，随机森林模型在向量的预测上更为公平（如N/S），而准确率的略微下降可能是因为搜索的精度导致的（由于运算量的巨大，无法进行和决策树一样更为广阔、深层的搜索），可以看到后两个向量的最优深度还可以更深，前两个向量的最优特征值数量也还在震荡

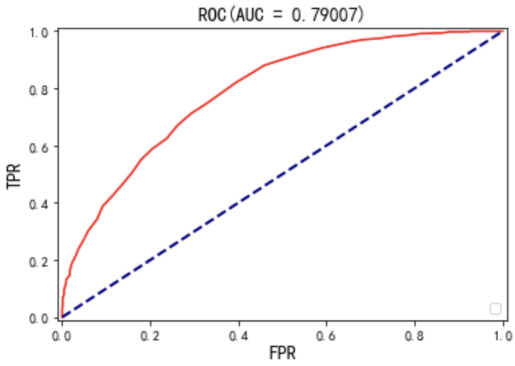


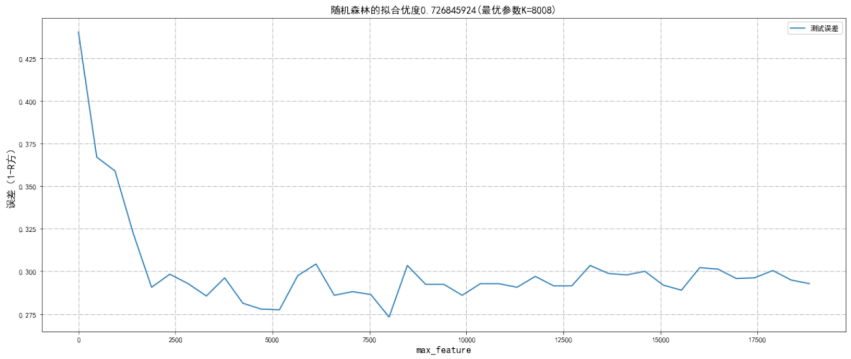


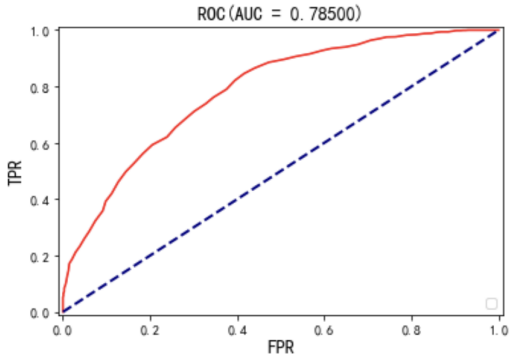










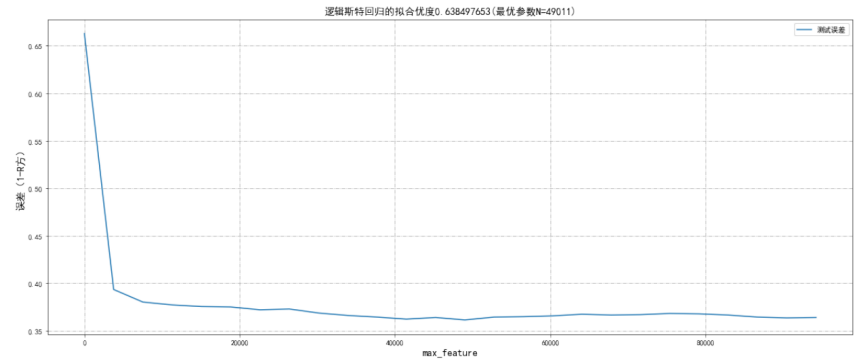


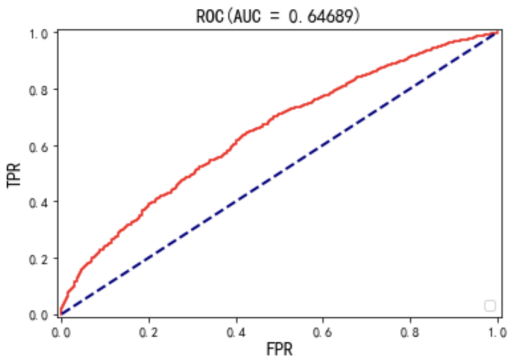
0.7504268032437047

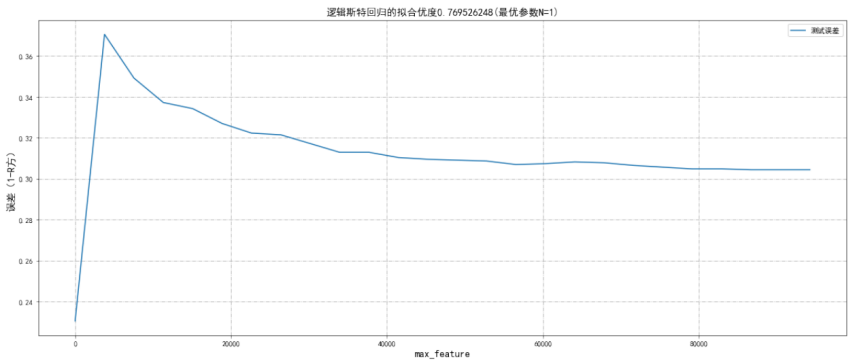
逻辑斯特回归

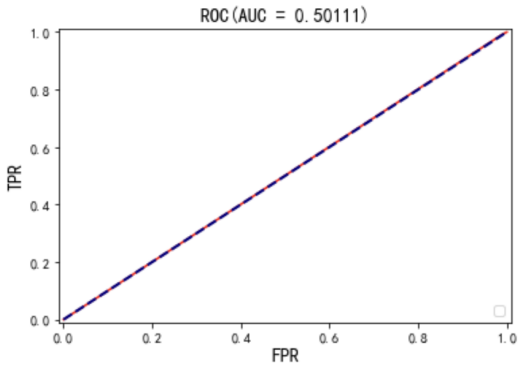
从上到下依次为I/E, N/S, F/T, P/J在对Tf-Idf最优特征值数量进行搜索后，输出的最优结果，以及其ROC曲线

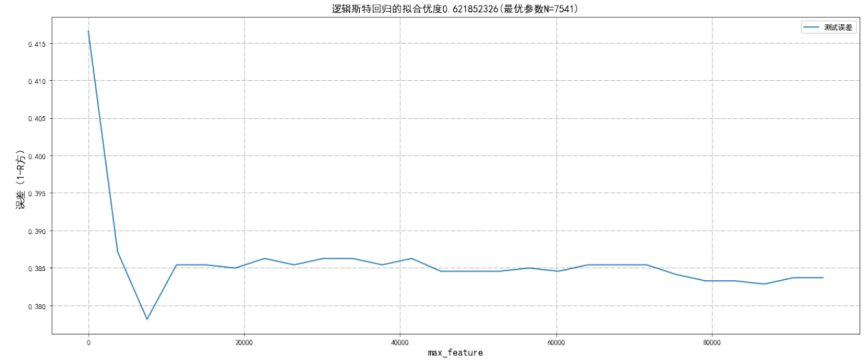
可以看出逻辑斯特回归效果很差，有的指标完全偏向了一方

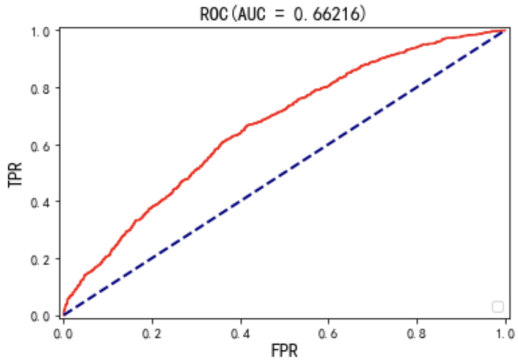


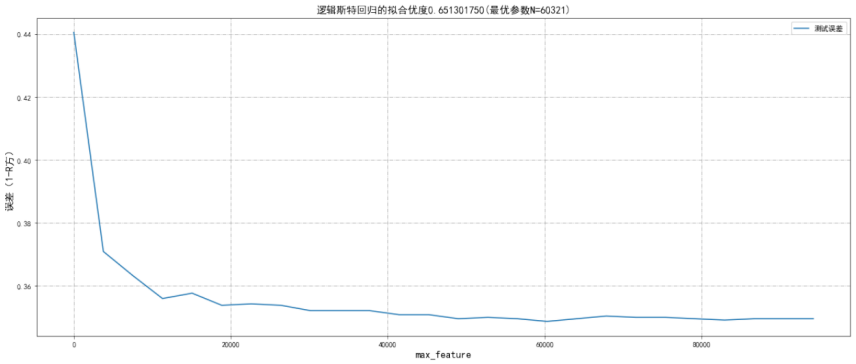


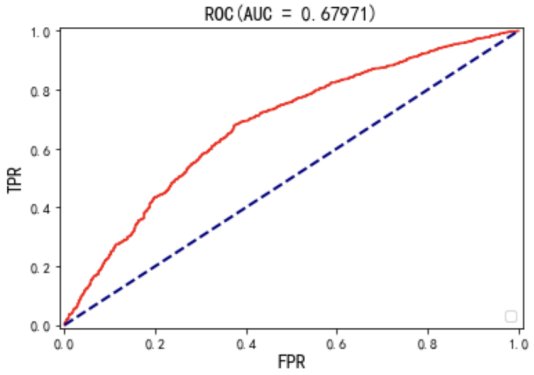












0.6702944942381562

支持向量分类机

线性支持向量分类机

```
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%
```

广义线性支持向量分类机

```
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%
```

总结支持向量分类机

分析实验结果，讨论评论文本与MBTI类型之间的关系

神经网络

由BERT模型生成的句向量作为评论的高维表示，使用双向上下文来理解单词的含义，能在一定程度上捕捉到句子中词与词之间的依赖关系和语义关联。同时由于其高度抽象化的表示，使得我们很难通过具体的词语对应特征表现，但是通过实验我们发现相比于（10，384）的评论矩阵（仅使用用户前10条评论生成的句向量矩阵），输入大小为（100，384）的评论矩阵在神经网络中具有更好的分类效果。这表明评论矩阵可以一定程度上反映用户的人格特征，特别的，具有更加丰富语义特征的评论矩阵能够更好的反映出用户的人格特征。

探讨发现和可能的解释

VI. 讨论

总结实验结果的主要发现

文本编码模型讨论

1. 综合Word2vec词向量在文本分类预测模型的应用来看，在N/S人格维度的分类任务上准确率最高，在P/J人格维度的分类任务上准确率最低，可以看出词向量编码考虑了词的上下文以及逻辑顺序，更突出了网络评论中的一个人的感知方式型人格，弱化了一个人的生活方式型人格的特征。
2.

分类预测模型讨论

神经网络

通过比较多个单目标模型与多目标模型的综合分类结果，单目标模型在各自目标的分类情况均优于多目标模型，这是由于在单目标任务中，神经网络可以专注于学习单个目标的特征和模式，然而，在多目标分类任务中，神经网络可能难以同时捕捉到多个目标的特征和关系，同时神经网络在学习特征表示时可能面临特征共享和特征冲突的问题，使得模型难以准确地区分和捕捉每个目标的特征。但是综合准确率上，多目标模型准确率更高，这可能反映出这四个不同的目标人各维度之间可能存在复杂的关联性和相互影响，而多目标模型能够更好地利用目标相关性、共享特征来进行参数的优化平衡。

同时在调参过程中我们发现，对于多目标模型，当提高性格取向维度损失函数的权重时，感知方式维度的预测效果会有小幅度的下降这可能是因为针对二者的优化目标可能存在冲突。优化其中一个目标的同时可能会对另一个目标产生负面影响。这可能需要更深层次的特征提取才能够解决。

决策树、神经网络、逻辑斯特回归

写在输出数据那里了

支持向量机分类

讨论实验的局限性和可能的改进方向

VII. 结论

总结实验的主要贡献和结果

强调研究的重要性和潜在应用领域

VIII. 参考文献

引用在实验报告中提及的相关文献和资源