编号：

哈尔滨工业大学
“大学生创新创业训练计划”
创新训练项目申请书

项目名称： 基于文本分类的诈骗案情分析方法研究

拟申请经费 650 **元**

执行时间： ２０２２ **年** ９ **月至** ２０２３ **年** ９ **月**

负 责 人： 徐浩铭 **学 号 ：** 2021112905

联系电话： 17341432003 **电子邮箱：** 978545377@qq.com

院 系 及 专 业： 未来技术学院人工智能领域方向

指导教师： 孙承杰 **职 称 ：** 副教授

联系电话： **电子邮箱：** sunchengjie@hit.edu.cn

院 系 及 专 业： 计算机科学与技术学院

哈尔滨工业大学本科生院制

填表日期： 2022 年 9 月 8 日

一、课题组成员：（包括项目负责人、按顺序）

姓名	性别	所在院	年级	学号	身份证号	本人签字
徐浩铭	男	未来技术学院	2021	2021112905	511923200304134018	
杜佳兴	男	未来技术学院	2021	2021110962	140121200308027539	
樊宇宇	女	未来技术学院	2021	2021111004	142701200404091222	
王雅斌	男	计算学部	2021	2021110963	140110200301302535	

二、项目简介（限 500 字以内）：

本项目拟基于文本分类相关算法实现一个操作简便，结果准确的诈骗案情分析交互式机器人。近几年，国内网络诈骗犯罪活动猖獗，案件数量逐年上升，给社会安定带来了极大威胁，也给警务人员带来沉重的工作压力。当前，自然语言处理(NLP)正处于快速发展的阶段，而文本分类是 NLP 中的一个经典问题，广泛应用于情感分析、垃圾过滤等领域。因此，本项目受此启发，将运用文本分类任务来识别诈骗案件，辨别其属于哪种类型的诈骗案件。

项目的研究内容主要分为五部分，第一部分构建语料库，该部分由公安局提供相关数据。第二部分，进行数据的预处理(分词处理、去噪、向量化)。第三部分，选取多种模型作为文本分类器，在机器学习中选择一些经典的机器学期算法，如 SVM，贝叶斯，KNN 等达到学习目的；除此还将选择深度学习中的 FastText 等合适的模型达到理想的精确率，从而实现项目的有效性。第四部分，结合 k 折交叉验证和网格搜索调参优化。最后一部分实现项目的可视化，用 flask 来实现网站的开发，同时最好将分类机器人嵌入网站之中。

本项目难点主要在于如何达到理想的精确率和召回率，从而实现项目的有效性。

三、申请基础（限 300 字以内）：

团队四位成员均有一定的基础编程知识与素养，对 python 有一定的了解，并且在大一年度项目和相关课程中对深度学习方面的知识有所了解。在之后我们会通过阅读论文、查看书籍、请教老师等方式来强化项目所需的知识与技能。

团队成员对自然语言处理均有一定的兴趣，这更利于我们去学习相关方面的新知识。

团队成员均参加并顺利完成了大一年度项目，所在小组获得一等、二等奖。有一定的合作经验。

1、立项背景

近几年，随着科技和互联网的快速发展，给人们生活带来巨大便利的同时也带来了电信诈骗的迅速蔓延，诈骗形式和诈骗手段的变化莫测给公安机关侦查取证工作带来很大挑战。据统计，借助电信网络技术诈骗犯罪的案件已经接近所有刑事案件的一半，而且这种状况还在以加速度的趋势演进。但是目前我们对各种电信诈骗方式、手段少有个性化的分析研究，尤其缺少精准的以数据为基础的类别性研究，将各个诈骗案件进行分类，可以更有了利于公安机关研究电信诈骗之间的交叉、区别和现状[1]。

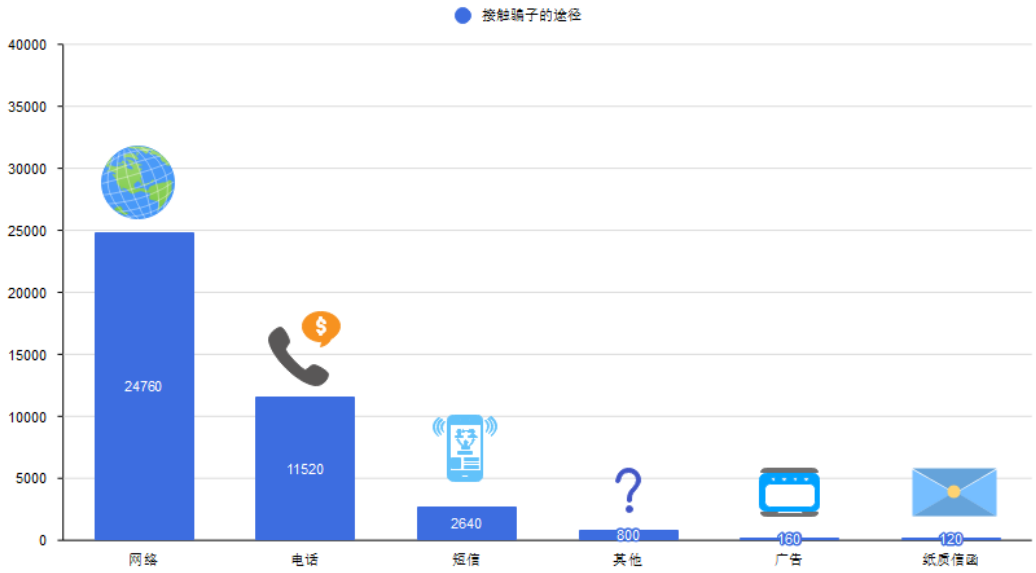

图 1 接触骗子的途径

人工智能技术的发展为诈骗案情的分析带来了新的机遇，通过自然语言处理中的文本分类技术，可以将文本集进行自动分类标记，有利于处理海量的数据，提高数据的利用率。本项目致力于构建一个以文本分类技术为基础的诈骗案情分析交互式机器人，对诈骗案情文本进行自动分类，实现对诈骗案情的精准分类，从而可以一定程度上减轻公安机关的工作压力，也能让人们第一时间知晓自己是否正处于骗局之中，减少人们的财产损失。

2、研究内容

2.1 文本分类

文本分类又称为自动文本分类，定义为利用计算机对文本及按照一定的分类体系和标准进行自动分类标记，通过一个已经被标注的训练文档集合，找到文档特征和文档类别之间的

内在关系模型，然后利用这种学习的得到的关系模型对新的文档进行类别判断[2]。其应用场景有新闻主题分类、情感分析、舆情分析、邮件过滤等等，其主要系统可以用下图来表示

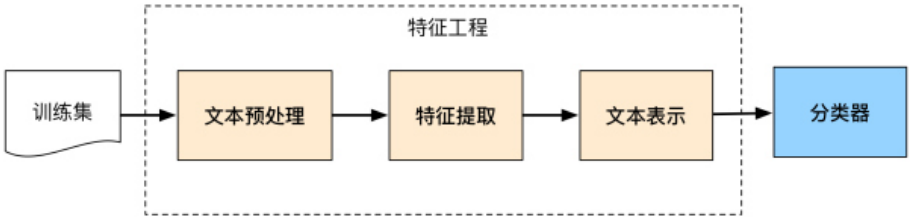

图 2 文本分类系统表示

2.1.1 训练集

数据集是存储于计算机中并可利用计算机进行检索、查询和分析的语言素材总体，其构建过程包括采集和预处理。数据集的采集规模、覆盖率和预处理决定其应用性能。近年来，随着深度学习及数据驱动自主学习高级特征技术的发展，大规模、高质量数据集的需求越来越大。数据集预处理不但要结合语言自身的特征，还要考虑其应用领域[3]。在本项目中我们将运用公安局提供的数据来作为训练集来进行训练。

2.1.2 文本预处理

文本预处理过程是在文本中提取关键词来表示文本的过程，中文的文本处理大致可分为文本分词和去停用词两个阶段。停用词的存在会增加文本数据处理的复杂度，还会降低文本的遍历速度，降低模型效率，去除停用词是指除去文本中常见但是没有明显语义的词语，例如助词、介词、语气词等。另一过程分词是中文文本处理中特有的且尤为重要的步骤,其质量高低严重影响后续文本处理与模型构建效果。目前中文分词技术根据实现方法可分为四类，分别是字典匹配、知识理解、统计信息、深度学习[4]。目前常用的中文分词工具有 Jieba、SnowNLP、PKuseg、HanLp 等。本项目目前考虑用 Jieba 或 HanLp 完成文本的分词处理。

Jieba 分词先采用基于字符串匹配算法，如果该单词存在于字典中，则拆分出该单词，否则就不对该单词进行分割；接着基于统计分词算法，其核心思想是在不同的文章中几个相连的字出现的频率越高，就说明这几个相连的字极大可能是一个词。Jieba 就是将这两种方法结合起来使用。其会形成一个层级结构，在这种结构下，每种分词的方法都可以找到一条从首字到末字的路径，在每个路径中，一个词语就是一个节点，词语和词语间的分隔符就是边。在进行切分文本时，首先根据前缀词典将所有的分词结果都切分出来，使用分词结果构建一个有向无环图，图中包含了所有分词的路径;接着再进行文本标注，得到最大的概率路径进而

得到最终的分词结果[5]。

HanLP 是由一系列模型预算法组成的工具包,结合深度神经网络的分布式自然语言处理，具有功能完善、性能高效、架构清晰等特点，支持中文分词、覆盖了词性标注、命名实体识别、词法分析、句法分析、文本分析和情感分析等常用任务，提供了丰富的 API，是 GitHub 最受欢迎、用户量最大、社区活跃度最高的自然语言处理技术。目前，HanLP 分词器已经被广泛应用于各种平台，有大量开源的各种作者研发的插件与拓展功能，并且被包装或移植到 Python、C、R、Java 等语言上去[6]。

2.1.3 文本表示

文本表示是自然语言处理的基础工作，他的处理结果会对整个自然语言处理网络造成极大的影响。用通俗的话说就是将文本转化成一系列能够表达文本语义的向量。在本项目中将使用分布式文本表示法的代表模型 Word2Vec。

Word2Vec 是一种词嵌入模型，是由谷歌团队在 2013 年提出的一个基于神经网络的语言模型，其支持对百万级的语料库进行高效训练；同时其训练的词向量语义表征性强，并且能够度量单词之间的词义相似度。其包含两种训练模型分别是 CBOW 和 Skip-gram。这两种是由 Mikolov 在 2013 年提出的基于词向量的无监督文本表示方法[7]。

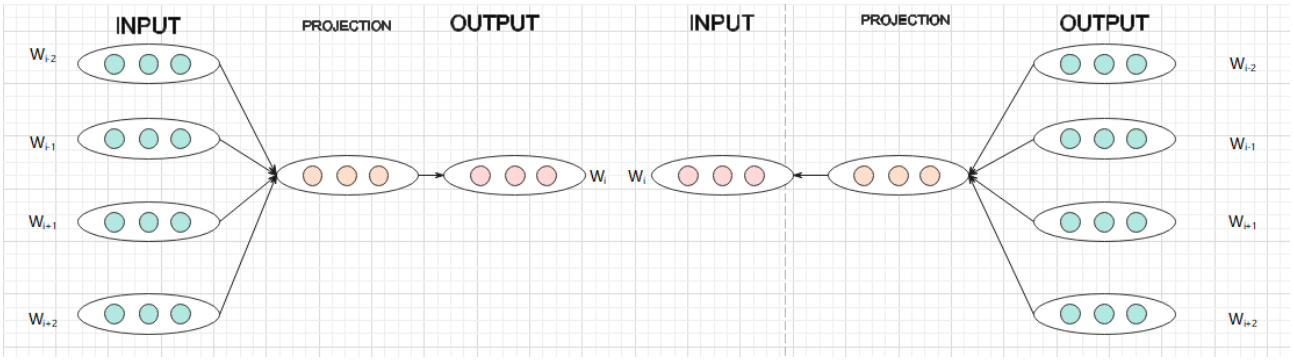

图 3 CBOW(左侧)模型与 Skip-gram(右侧)模型

CBOW 和 Skip- gram 模型。如果是用一个词语作为输入，来预测它周围的上下文，那这个模型叫做 Skip- gram 模型。如果拿一个词语的上下文作为输入，来预测这个词语本身，则是 CBOW 模型。由于 CBOW 和 Skip-Gram 模型的原理相似，所以主要介绍 CBOW 的映射原理。

CBOW 主要分为输入层、隐藏层和输出层。如果将当前滑动窗口范围内的 k 个词作为上下文信息，通过滑动窗口方式预测中心词时，中心词概率公式如下所示:

$$\mathrm{P}\left(W_t \middle| W_{t-k}, W_{t-(k-1)}, \ldots, W_{t+1}, W_{t+2}, \ldots, W_{t+k}\right) = \mathrm{P}(W_t|\mathrm{context})$$

其首先随机生成一个包含文本中所有单词的词向量矩阵，其中矩阵的每一行都是一个单词的向量表示;然后从词向量矩阵中提取中心词上下文单词的词向量，并计算这些词向量的平均值;再对平均值向量做逻辑回归训练，选用 softmax 作为激活函数;训练结束后，比较所得概率与中心词的概率向量是否接近[8]。

Skip-gram[9]与 CBOW 的操作过程相反，它主要是根据已知的中心词信息来预测未知的上下文信息。

其算法过程如下:首先随机生成一个包含文本中所有单词的词向量矩阵，每一行是一个单词的向量;然后从该文本中随机选取一个单词并对这个单词的词向量进行提取;再选择 softmax 作为激活函数并对提取的词向量进行逻辑回归训练;训练结束后，比较得到的概率与上下文的概率向量是否接近。Skip-gram 属于非监督学习，在数据集比较大的情况下，结果更精确。

2.1.4 分类器

(1)机器学习模型

将文本表示为模型可以处理的向量数据后，就可以使用机器学习模型来进行处理，常用的模型有 SVM，贝叶斯，KNN 等，我们将采用贝叶斯等算法来进行学习。所以下面详细介绍贝叶斯的原理。

朴素贝叶斯是一种比较简单的模型，它是一种基于统计的算法。该分类算法把最高概率的类别作为该数据类别，选择最高概率决策是朴素贝叶斯分类算法的核心思想。假设有 n 个类别$(C_1, C_2, \cdots, C_n)$，x 为待分类数据项，需要计算 $P(C_i \mid x)$, 即 x 属于每个类别的概率, 利用贝叶斯定理，公式如下:

$$P(c_i|x) = \frac{p(x|c_i)p(c_i)}{p(x)}$$

取其中最大概率类别，即为数据项 x 所属类别。

朴素贝叶斯分类方法，是通过先验概率和可能性函数计算出后验概率，从而将待分类数据项进行分类。贝叶斯分类方法对于多分类问题有不错的分类效果，而其基于统计的理论依据也很容易理解和实现[10]，所以我们将选择贝叶斯进行学习。但在实际应用场景中，朴素贝叶斯算法中的条件独立性假设不一定成立，为此我们针对这一问题对贝叶斯方法进行一定的改进。

(2)深度学习模型

应用深度学习解决大规模文本分类问题最重要的是解决文本表示，再利用 CNN/RNN 等网络结构自动获取特征表达能力，去掉繁杂的人工特征工程，端到端的解决问题。目前主要的有 FastText，TextCNN，BERT 等模型。我们将采用 TextCNN 模型来进行该任务。以下详细介绍 TextCNN 模型。

在 2014 年 Kim[11]等人首次将 CNN 引入自然语言处理领域，应用到文本分类任务中。卷积神经网络是一种带有卷积操作的前馈神经网络，用于文本分类的卷积神经网络 TextCNN 在 CNN 的基础，对输入层进行了一些改动，但在网络结构上与 CNN 没有任何变化。如图 4 所示为 TextCNN 的网络结构，整个网络包括三个部分，输入层，计算层和输出层。

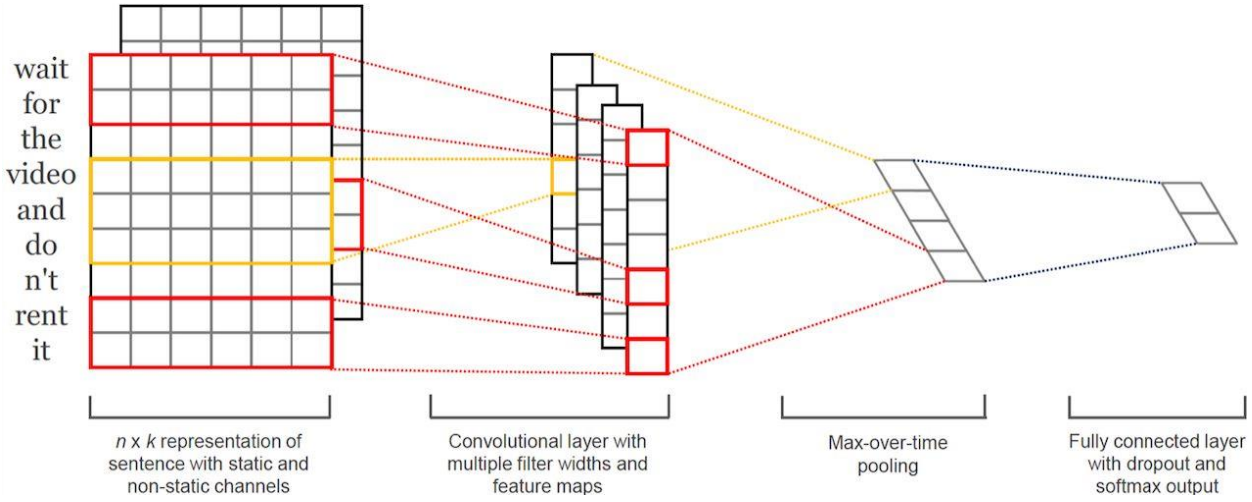

图 4 TextCNN 网络结构

输入层由 Word2vec 处理后的文本进行词嵌入，通过 K 维词嵌入后，一个词语数量为 N 的文本被编码为 N×K 的二维特征矩阵。由于句子长短不一，所以还需要进行归一化处理，长度不够 N 的用 0 进行填充，长度大于 N 的将过长的部分进行截断丢弃。

TextCNN 的计算层由一个卷积层，一个池化层以及全连接层构成。卷积计算之后拼接三个不同维度的特征图，再使用最大池化进行池化。接着再使用全连接层对特征进行降维。输出层使用 softmax 函数得到每个文本的类别[12]。TextCNN 在卷积层的操作属于一个线性操作，但是，文本数据信息复杂多变，并不是线性的，单一线性函数是无法解决非线性问题的，所以需要在模型结构中增添一层激活层，使得模型能够解决像多分类这种非线性问题。模型常用的激活函数有如下几种:

(1) Sigmoid 函数:输出范围在(0,1)之间，优化稳定,求导简单，但是容易出现梯度消失的问题。

$$Sigmoid(x) = \frac{1}{1+e^{-x}}$$

(2)tanh 函数:要比 Sigmoid 激活函数的收敛速度快，但是由于其需要繁琐的幂运算,计算成本较高,同时也存在梯度消失的问题。

$$\tanh(x) = \frac{e^{x} — e^{-x}}{e^{x} + e^{-x}}$$

(3)ReLU 函数

ReLU 激活函数应用十分广泛，因为其不需要进行幂运算，所以运算速度也较快,相对于上面两种激活函数而言,ReLU 函数也可以极大改善梯度消失问题，同时由于 ReLU 函数会失效一部分神经元，这样的特点可以减少模型发生过拟合[13]。

TextCNN 作为用于文本分类的神经网络，网络层数不多，结构简单，计算量小，网络收敛快。作为局部寻优的模型，使用预训练效果好的词向量，TextCNN 可以在短文本分类取得很好的分类效果，但若文本序列较长的话，其特征提取能力受限，导致分类效果不佳。

2.2 网页框架

基于上述文本分类的模型构建与优化，为了实现项目的可视化。考虑搭建一个网页，同时将分类机器人嵌入，部署上服务器。目前有很多比较知名的 web 框架，分别是 Diango、Tornado、Flask。其中 Django 是市场占有率极高的框架，适合大项目，官方文档齐全；Tornado 的异步高性能框架，包含许多底层细节，少而精； Flask 微框架，轻量级，扩展插件较多。在本项目中我们将使用 Flask 来搭建网页。

使用 Flask 是因为其是轻量级的微框架，扩展插件比较多，其超高的扩展性和小而精的核心本身。此外选用 Flask 学习所需的时间较短，上手快，所以将使用 Flask 框架。

Flask 是由 python 语言编写的 web 框架，其上手简单、开发便捷，当了解到其基本功能，并且进行基础开发之后，就可以比较轻松的阅读它的源代码。其虽然是小型框架，但由于其满足可延展性，其核心只有一系列的基本服务，而其他服务都可以通过 PyPI 进行下载或者自己编写，从而避免了框架的沉重繁杂。

Flask 的工作流程为:在用户访问 URL 时，通过 WSGI(Python Web Server Gateway Interface)协议将请求信息转换为服务器处理的相应接口格式，调用服务器的相应函数生成返回信息，经过 WSGI 协议转换格式，最后传递至前端界面展示该信息[14]。工作过程如图所示：

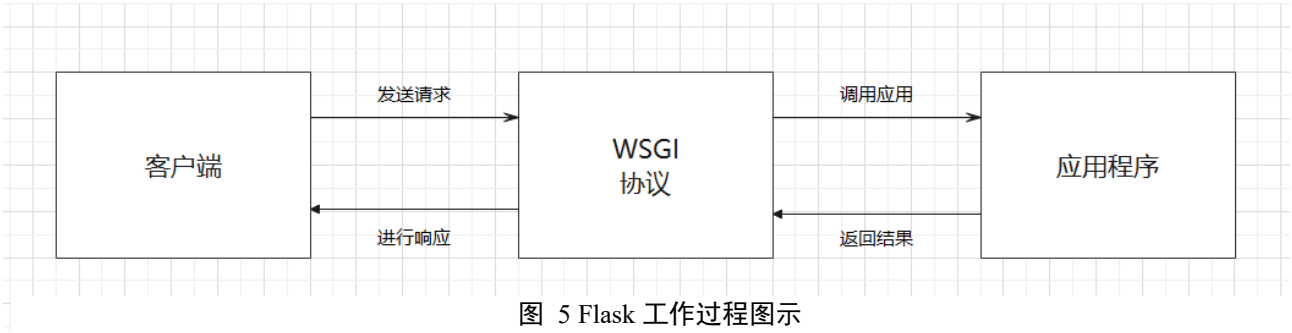

图 5 Flask 工作过程图示

2.3 小程序构建

基于上述文本分类的模型构建与优化，设计一款简单的针对诈骗文本分类的反诈小程序，包括最为基础的网页框架的搭建，前后端的对接，服务器的部署与进阶一步的分类机器人的嵌入等。通过这些研究，提高人机交互性，不断丰富用户的使用体验。

3、预期目标

3.1 中期目标

（1）完成前序知识的学习与准备(包括 python 的使用，机器学习模型，深度学习算法与自然语言处理基础知识的学习)。

（2）良好的完成数据预处理(包括分词处理，去噪，向量化等过程)，以求最大程度上减少后期计算的内存开销和计算误差，为后期的模型运用与准确率提升打好数据基础。

（3）初步简单实现项目完整的流程。

3.2 结题目标

（1）继续深入学习项目相关内容 。

（2）优化模型，提高模型准确率与召回率。

（3）完成较好的可视化，搭建项目相关的网页框架。

（4）完善项目的各个流程，撰写好一篇报告。

4、特色与创新

(1)在完成模型的初步实现后，使用 k 折交叉验证和网格搜索进行调参优化。网格搜索能够通过穷举法得到最优的超参数组合，然而，本项目中关于诈骗网站的数据量较小，在网格搜索的过程中，验证集中并没有足够多的数据，不能代表总体数据的特征，因此调参过程具有极大的偶然性，会影响到最终模型的准确率。

针对这个问题，本项目小组考虑采用 k 折交叉验证法，通过充分利用数据来避免这种偶然性，选择超参数的最优组合。

同时，在使用 k 折交叉验证法的过程中，由于每一个数据都有可能充当训练集，验证集或者测试集，因此所有数据都会参与到训练和测试的过程中，从而能够有效解决欠拟合和过拟合的问题。不仅如此，使用 k 折交叉验证还能评估各个模型的质量，选取给定数据集上的最优模型。因此，结合网格搜索与 k 折交叉验证的方法，能够有效提升文本分类模型的准确率，增强本项目中文本分类器的效果。

(2)项目计划构建多个基本的分类模型，基于此计划搭建多个文本分类器，后面通过一些算法例如 Bagging、Boosting 和 Stacking 等进行模型结果的综合，通过综合多个分类结果，能够在很大程度上提升结果的准确率。

(3)项目在完成了模型建立与优化调整的基础上，搭建前端网页框架，考虑嵌入分类机器人，部署服务器，更进一步完成一个小程序的初步构建，推动文本分类功能进一步完善，人机交互性进一步增强，从而能够从创新的小切入口逐渐过渡到初步的创业雏形。通过这些可视化过程，本项目不仅仅停留在理论上的分类模型建立，更能够进一步走向反诈的实践，更大的应用与实践空间也能够反向拉动模型的进一步优化，实现理论与实践的结合与良性循环。

5、实施方案

按照项目实施的时间先后顺序，本项目可以大致划分为五个步骤：完成前序知识的学习和准备—数据预处理---模型—优化—可视化和交互

第一步是小组成员完成前序知识的学习和准备，学习的内容有 python 语言的学习，自然语言处理，相关软件的使用。主要的学习资源有：《python 程序语言设计》MOOC，自然语言处理入门（何晗），程序员记笔记软件 Typora。这一步骤准备以自学 MOOC 和小组讨论的方式进行，具体实施方案如下：在项目最开始的两个月内，小组成员在周一到周五按照进度自学所需内容并做笔记，笔记上记录与项目重点相关的内容，复杂的知识以及学习过程中遇到的困难，同时在周末可以进行小组讨论报告学习进度，提出这一周学习的困境，进行解决或者请教老师解决。这样小组成员之间可以起到相互监督的作用，能够保证每个小组成员可以按部就班完成学习内容。

第二步数据预处理又可分为分词处理，去噪，向量化。这一部分小组成员可以通过自习自然语言处理入门，相关文献，小组讨论以及老师的指导来学习。目前考虑用 Jieba 或者 HanLp 完成分词处理，经过学习与指导后确定用某一种方法实现。去噪可以使用停用词词典文件：

[data/dictionnary/stopwords. txt]，该词典收录了常见的中英文无意义词汇(不含敏感词)，每行一个词。向量化可以用词袋模型 CountVectorizer 或者分布式表示如 Word2vec。

第三步模型，我们会综合选取合适的模型。对于机器学习，由于我们是初学，我们将首先采取经典的机器学习算法，如 SVM，贝叶斯，KNN 等达到学习和练手的目的。对于深度学习，采取更加复杂有效的模型，综合选取合适的模型如 FaxtText，TextCNN，BERT 并且最后达到理想的精确率和召回率，实现项目的有效性。

第四步优化，会结合 k 折交叉验证和网格搜索进行调参。。

最后一步可视化和交互，可以建立一个网页框架，考虑用 flask，同时最好将分类机器人嵌入，绘图模块用 matplotlib 实现。

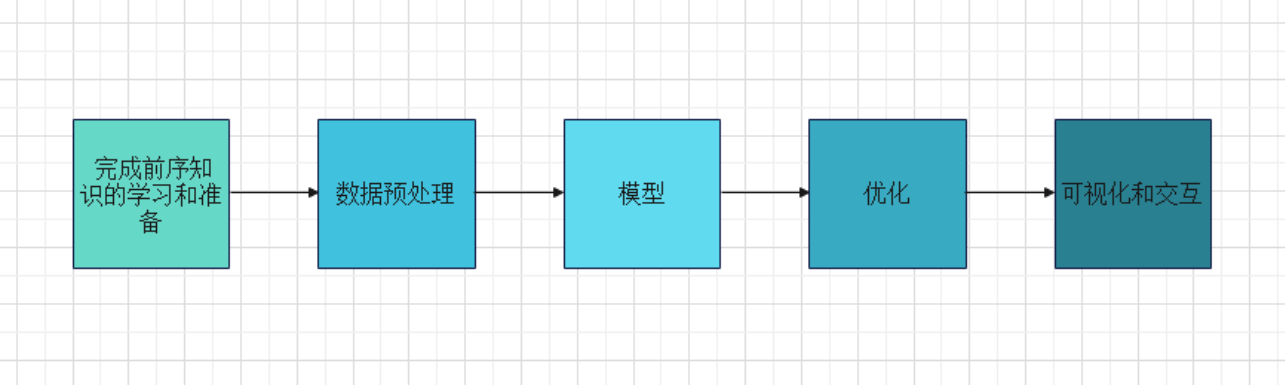

图 6 实施方案步骤图

6、进度安排

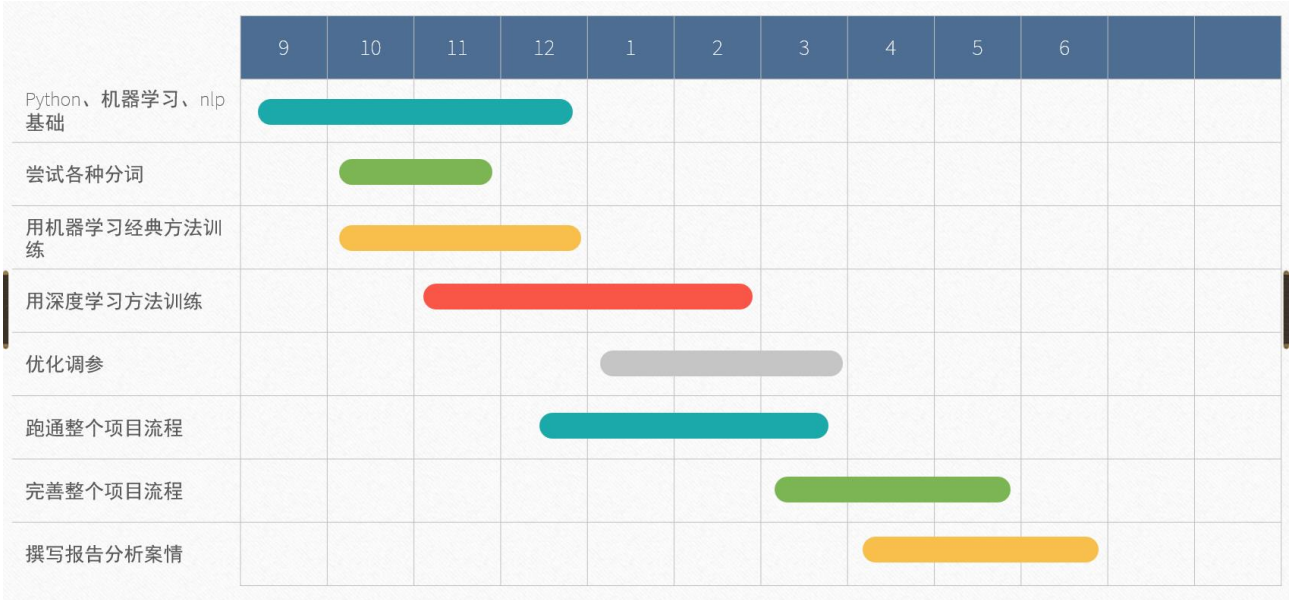

图 7 进度安排甘特图

7、经费预算

预算类别	主要用途	预算金额（元）
书籍费	购买书籍等	200
资料打印	打印所需的资料	50
服务器租聘	将服务器部署到网站上	400

8、参考文献

[1]熊春海. 电信网络诈骗犯罪的现状及治理完善路径[D]. 广西师范大学, 2022.DOI: 10.27036/d.cnki.ggxsu.2022.000940.

[2]Asad Abdi,Siti Mariyam Shamsuddin,Shafaatunnur Hasan,Jalil Piran. Deep learning-based sentiment classification of evaluative text based on Multi-feature fusion[J]. Information Processing and Management,2019,56(4).

[3]道吉扎西. 基于 TWC__CNN 的藏文文本分类技术研究[D]. 青海师范大学, 2022. DOI: 10. 27778/d. cnki. gqhzy. 2022. 000588.

[4]房京珂. 基于 NLP 的微博情感分析研究[D]. 中央民族大学, 2021. DOI: 10. 27667/d. cnki. gzymu. 2021. 000315.

[5] Ding Y, Teng F, Zhang P, et al. Research on Text Information Mining Technology of Substation Inspection Based on Improved Jieba[C]//2021 International Conference on Wireless Communications and Smart Grid (IcwCsG). IEEE, 2021: 561-564.

[6]王璐. 基于知识图谱的健康膳食知识智能问答系统[D]. 兰州大学, 2020. DOI: 10. 27204/d. cnki. glzhu. 2020. 001921.

[7]Tomas Mikolov, Kai Chen 0010, Greg Corrado, Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space[J]. CoRR, 2013, abs/1301. 3781.

[8]崔文艳. 基于时间卷积网络的商品标题文本分类方法研究[D]. 天津师范大学, 2022. DOI: 10. 27363/d. cnki. gtsfu. 2022. 001116.

[9]黄聪. 基于词向量的标签语义推荐算法研究[D]. 广东工业大学, 2015.

[10]于敏. 基于机器学习的文本分类方法研究[D]. 江南大学, 2021. DOI: 10. 27169/d. cnki.gwqgu. 2021. 000150.

[11]Yoon Kim. Convolutional Neural Networks for Sentence Classification.[J]. CoRR, 2014, abs/1408. 5882.

[12] 曾芳. 基于混合卷积的文本分类算法研究[D]. 西南科技大学, 2022. DOI: 10. 27415/d. cnki. gxngc. 2022. 000957.

[13] 郭书武. 基于深度学习的教材德目分类方法研究[D]. 上海师范大学, 2022. DOI: 10. 27312/d. cnki. gshsu. 2022. 001952.

[14] 刘嘉伟. 基于 FLASK 的校园智能停车系统的构建[D]. 吉林大学, 2021. DOI: 10. 27162/d. cnki. gjlin. 2021. 001834.

四、评审情况：

指导教师意见：   指导教师签名：　　年　　月　　日