编号：2022F0316

哈尔滨工业大学
大学生创新创业训练计划项目验收书

项目名称： 基于文本分类的诈骗识别

项目级别： 省级 **（国家级、省级、校级）**

执行时间： 2022 年 9 月至 2023 年 10 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： 13633615141 **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

填表日期： 2023 年 10 月 6 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	徐浩铭
杜佳兴	男	未来技术学院	2021	2021110962	140121200308027539	杜佳兴
樊宇宇	女	未来技术学院	2021	202111004	142701200404091222	樊宇宇
王雅斌	男	计算学部	2021	2021110963	140110200301302535	王雅斌

二、指导教师意见：

项目完成了预定的目标，利用多种深度学习模型实现了一个准确率较高的诈骗案件描述文本分类系统，并完成了一个可视化效果较好的演示系统。项目组成员在完成项目的过程中掌握了很多最新的深度学习模型和大规模语言模型的知识；提高了分析问题解决问题的能力；养成了良好的团队合作精神。同意参加结题答辩。

签　　名：

2023　年 10 月　8 日

三、学院专家组意见：

组长签名：　　（ 盖　章 ）

年　　月　　日

四、项目成果：

（一）申请专利情况：

序号	专利名称	发明人	专利申请号	备注

注：请将专利申请书的电子版作为附件报送。

（二）发表论文情况：

序号	论文题目	作者	刊物名及期号	备注

注：请将所发表论文及当期刊物封皮、目录的电子版作为附件报送。

（三）其它成果（软件、模型、图纸或作品等）：

序号	名称	说明
1	交互式诈骗文本分类网页	基于 flask 框架和 MySQL 数据库，构建了一个集成各模型，包含各类别分析的前端交互网页。

五、项目研究结题报告

1、课题研究目的

本项目拟基于电信网络诈骗案件分类任务，实现一个操作简便，结果准确的诈骗案情分析系统。近年来，国内网络诈骗犯罪活动猖獗，案件数量不断上升，给社会带来了极大威胁[1]。同时，警务人员也面临着沉重的工作压力。但目前对各种电信诈骗手段和方式的个性化分析研究甚少，尤其缺乏以数据为基础的精准分类研究。因此，将各种诈骗案件进行精确分类尤为重要，对于个人来说，可以及时警醒，避免遭受潜在的欺诈；对社会来说可以更好地保护公众的合法权益和财产安全，这有助于社会建立良好的信任基础，增强社区的安全感和社会稳定性；对相关机关来说，有助于执法机构及时介入并采取适当的法律行动。这有助于制止诈骗活动，维护社会的法律秩序。同时通过对诈骗案件的分类和分析，可以更深入地研究犯罪模式、诈骗者行为、社会脆弱点等。

2、课题背景

社会上的诈骗情况呈现多样化、复杂化、技术化和全球化的趋势。网络技术的发展和普及使得网络诈骗成为主要的犯罪手段之一，诈骗分子利用数据获取的便利，衍生出了许多实施诈骗的套路，同时传统的诈骗方式仍然存在且翻新不断。图 1 和图 2 是对本年上半年手机诈骗的一些情况，可以看出诈骗的年龄反而是年轻化，话题是新潮化，所以当下做好反诈骗尤其重要。

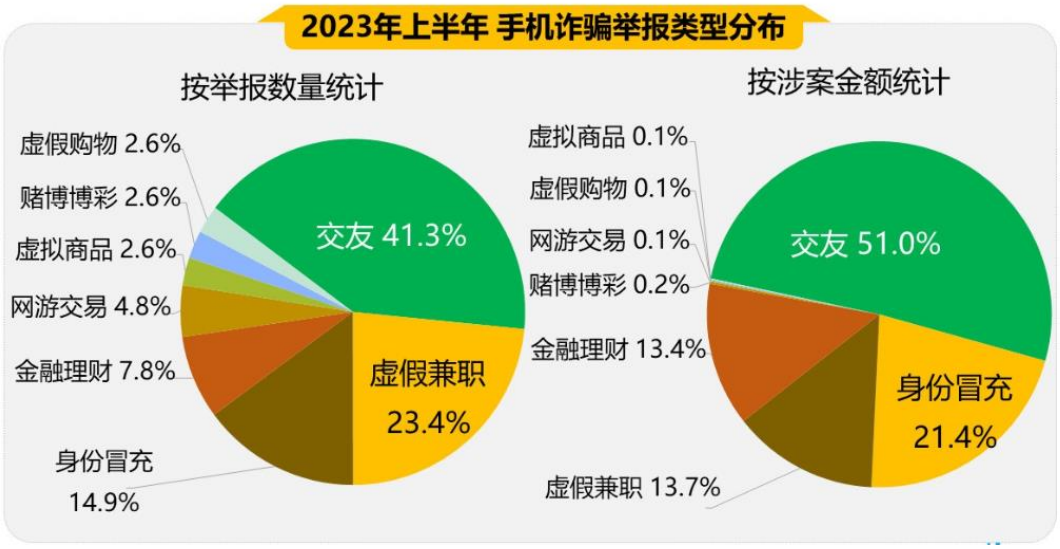

图 1 手机诈骗举报类型分布

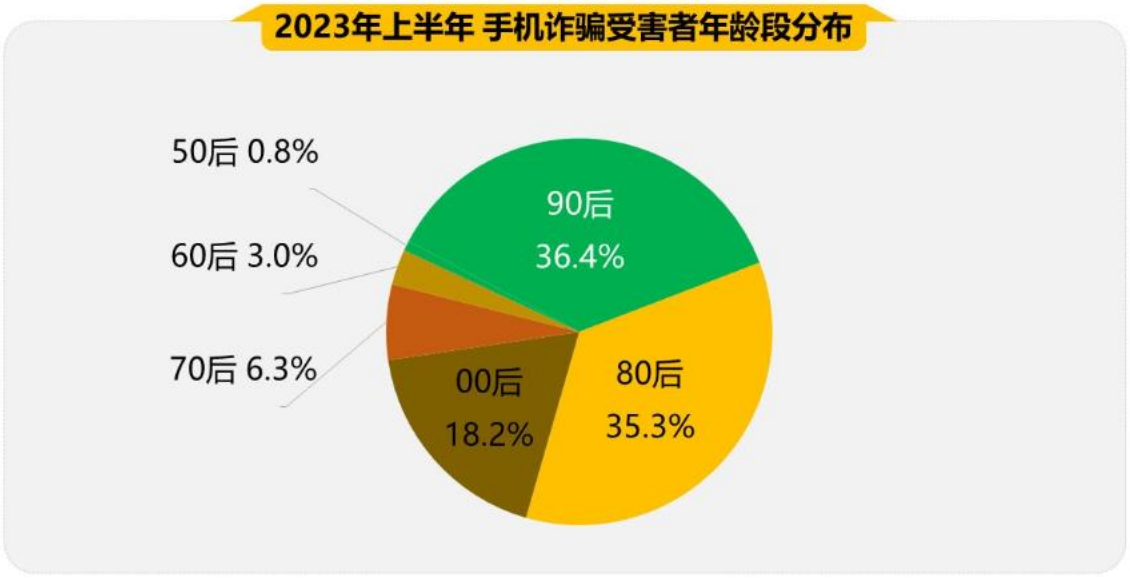

图 2 手机诈骗受害者年龄段分布

另外，全国检察机关今年一季度批捕诈骗犯罪 10923 人，在所有罪名中位列第 2 位，提起公诉 18146 人，在所有罪名中位列第 5 位。此外，组织化、规模化特征明显。检察机关起诉的诈骗案件中，五成以上为共同犯罪。诈骗团伙组织严密、分工明确，有的以所谓“合法公司”为掩护，租用高档写字楼，设立多个部门或岗位，利用网络平台进行宣传、招聘，采用企业运作模式管理。电信诈骗案件的侦办工作往往以受害人报警为触发点，公安机关传统的侦办方式更加侧重于对单一或单系列案件的研判，缺乏对一定时期内此类案件整体发案情况的掌握。这种犯罪研究方式难以从宏观层面发现电信诈骗案件的整体规律。加之电信网络诈骗案件作案手段种类多、变化快，也加大了提前预防的难度，使得公安机关处于疲于应付的局面，亟待探索一种新的案件研判预防思路 。

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，我们可以将文本集进行自动分类标记，从而更好地处理海量数据，提高数据的利用率。因此，本项目将为公安机关提供一个更高效、更精准的技术工具，以应对电信诈骗犯罪的挑战。

3、课题研究主要内容

3.1 前期工作及项目流程

3.1.1 前序知识的学习和准备

在项目中期，我们小组基本完成了一个完整的文本分类系统的搭建并且实现了初步的可视化与交互。

3.1.2 项目流程

图 3 项目流程示意图

如上图 3 所示，本项目流程可以大致划分为五个部分：完成前序知识的学习和准备 - 数据预处理 - 模型 - 模型校准与模型投票 - 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，具体学习的内容参见 3.1.1 前序知识的学习和准备部分。

第二步数据预处理我们进行了各类别数据分布的分析，数据清洗以及词频分析，并且在中期的基础上，我们增加了数据库来更高效的管理我们的数据。数据分布分析这一部分，我们对 12 个诈骗类别的数据量进行了统计与可视化分析。数据清洗这一部分中，我们对诈骗类型进行编码，去除了数据的敏感信息，并且对数据进行了分词[3]，去噪，向量化处理。词频分析部分我们对各个诈骗类型进行了词频分析与可视化。

第三步模型，我们综合选取了合适的模型。我们首先采取经典的机器学习算法，如 SVM，贝叶斯，KNN 等达到初步的分类效果。对于深度学习，在项目中期结果的基础上，我们采取了更加复杂有效的模型，综合选取合适的模型如 FaxtText,TextCNN,BERT 等并且最后达到理想的精确率和召回率，实现项目的有效性。

第四步模型校准与模型投票。在这一步中，我们利用了一些常见的模型校准方法，例如 temperature 平滑，数据 label 平滑等提升了模型的准确度。并且利用模型投票对上述单一模型进行简单的集成。

最后一步可视化和交互，我们进一步优化了项目中期时搭建的可视化网页，实现内容分析和基本功能的展示，并且用户能够通过网页评估模型分类结果，进而提升了网页的交互性，同时能够利用每次分类任务的结果优化后端数据库，提升分类模型的准确性。

3.2 数据处理

3.2.1 数据处理

（1）数据采集

数据由公安部门反诈大数据平台导出，每一条数据包含案件文本和类别标注，其中案件文本内容为案情简述，即关于案件经过的描述性文本，具体示例可参考下面的处理结果示例部分。

（2）数据清洗

在这一部分，我们去除了原始数据中“其他类型诈骗”的类别，对剩余 12 个诈骗类型

进行了编号，具体编号以及每个诈骗类型对应的数据量如下表 1 所示。并分别去除了时间信息，一些各类别中笼统代表诈骗的普遍的高频词（这部分词我们认为不能作为分类的特征词）以及特殊字符和停用词。

表 1 个诈骗类型及其编码

类型	编码	数量
冒充电商物流客服类	0	11018
贷款、代办信用卡类	1	8883
虚假网络投资理财类	2	9469
冒充领导、熟人类	3	3525
冒充公检法及政府机关类	4	3651
网络游戏产品虚假交易类	5	1723
刷单返利类	6	28367
虚假征信类	7	6771
冒充军警购物类诈骗	8	873
虚假购物、服务类	9	5647
网黑案件	10	958
网络婚恋、交友类（非虚假网络投资理财类）	11	1324

（3）脱敏处理

为防止对受害者造成二次伤害以及防止诈骗信息产生二次传播，去除了数据中受害者的隐私信息及诈骗分子的不良信息，具体来说，包括案件文本中的姓名、出生日期、地址、涉案网址、各类社交账号以及银行卡号码等信息。

（4）后续处理

由于文本都很长，通过观察处理后文本小于 10 的数据包含信息很少，故去除文本小于 10 的数据。数据根据 8：1：1 划分 train ： test ： valid。

表 2 处理后的文本长度

平均长度	162.64
最大长度	1046
最小长度	10

文本长度可视化结果如下图 4 所示：

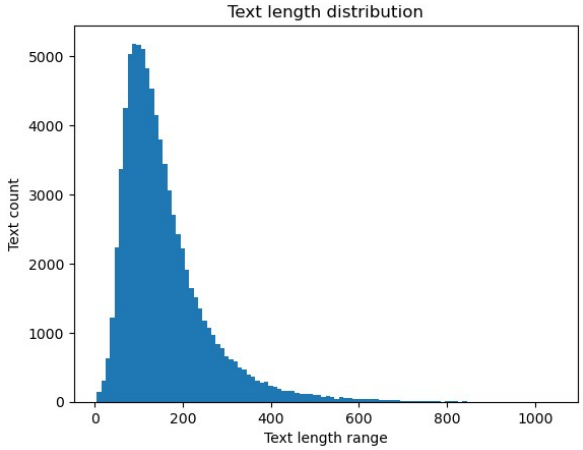

图 4 文本长度可视化结果

3.2.2 数据样例

原始数据样例：数据以 json 格式存储，每一条数据具有三个属性，分别为大类编号、简要案情、案件类别。

{
"大类编号": 0,
"简要案情": "2022 年 6 月 7 日，报警人（女，20 岁，人，身份证号,，兴盛社区，无业）报称：其接到自称是拼多多客服的电话，称其其在拼多多的衣服丢失（当事人无法提供商品信息），现可以进行赔付，对方以赔付需要刷流水的方式，按照对方的操作方法转账，转完账后发现被诈骗，总共被骗 17781 元。具体转账记录，2022 年 04 月 25 日通过名下银行卡转账 15781 元给银行卡:；转账 2000 元至的工商银行卡上，卡号:。对方的电话号码:,,,,，,，„„；报案人的农业银行卡号:。",
"案件类别": "冒充客服类"
},

处理后的数据样例：

（去掉时间信息，受害人信息）报称其接到自称是拼多多客服的电话称其其在拼多多的衣服丢失当事人无法提供商品信息现可以进行赔付对方以赔付需要刷流水的方式按照对方的操作方法转账转完账后发现被诈骗总共被骗（去掉不属于特征词的具体数字）元具体转账记录通过名下银行卡转账元给银行卡（脱敏处理）；转账元至的工商银行卡上卡号对方的电话号码（脱敏处理）；报案人的农业银行卡号（脱敏处理）

3.2.3 数据分布

我们对数据的 12 个类别的分布情况做了分析，得到的各类别数量如上表 1 所示。可视化结果如下图 5 所示：通过观察上述可视化结果，我们可以发现类别之间是不均衡的。

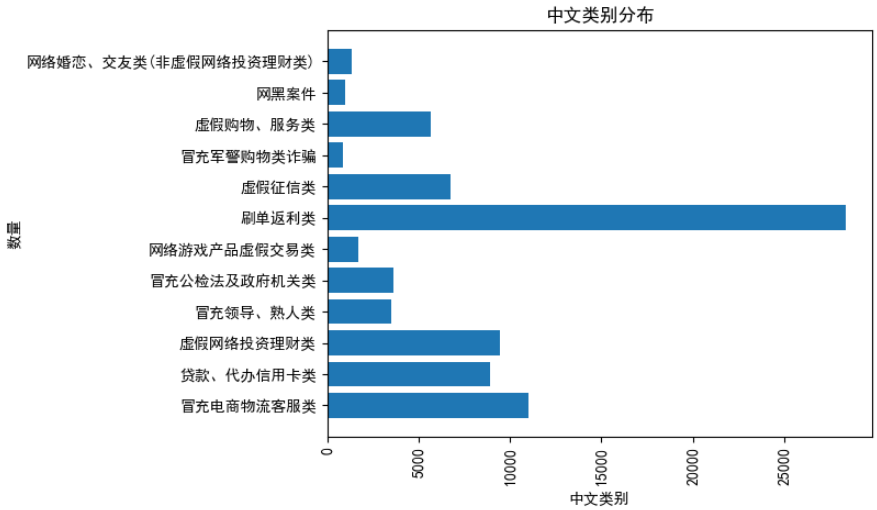

图 5 各诈骗类别数量可视化结果

3.2.4 分类别词频统计

我们小组采用文本计量方法对电信诈骗案件的案情描述部分分类别进行词频统计和计量分析。词频分析是一种内容分析方法，根据词频统计的规则，区分出高频词汇和低频词汇，并采用“可视化”的方法，通过词云等文本分析的工具，可以直观的凸显出具有代表性的高频词汇，便于从多个维度加以分析。一个单词或名词或短语出现频率越高，就越能体现出整篇文书的主题，高频词反映出了电信诈骗案件中的特点、诈骗话术常用词等。

我们对各个类别诈骗案件的案情描述进行了词频分析。例如，冒充领导、熟人这个诈骗类型我们得到的词云图如图 6 所示：

图 6 冒充领导、熟人词云图

在词云图中，我们可以发现“朋友”“儿子”“女儿”等词出现频率较高，说明在这个诈骗类型中，诈骗分子多次通过伪装成家人和朋友实施诈骗。而“QQ”，“微信”等高频词说明了常见的实施诈骗的平台，诈骗分子通过盗取熟人的 QQ 或者微信等线上平台的账号，冒充熟人给被害人发消息，要求被害人给自己转账从而实施诈骗。这说明，为了减少此类诈骗，公安机关应该提醒民众多注意“朋友”，“儿子”，“女儿”，“微信转账”等等上述提及的关键词，这些高频关键词也突出说明了冒充领导、熟人类诈骗的特点。

3.2.5 文本表示方法

（1） 离散表示

文本的离散表示有 One-Hot、BOW、 N-Gram 以及 TF-IDF 等。

One-Hot 编码通过一个由 0，1 元素组成的向量表示词语，其中只有一个维度是 1。这种表示方式形成的向量过于稀疏，且不同词语之间的 One-Hot 编码相互正交，很难理解词语之间的关系。

BOW(Bag-of-words)在 One-Hot 的基础上，统计词表中每个词在文本中出现的次数，将向量中的 1 更新为词频，即通过每个词语 One-Hot 编码相加得到一段文本的词袋模型表示。仍然没有解决词语之间缺乏联系的问题。

N-Gram 利用马尔可夫链的思想，将每个词出现的频率与前 N-1 个词联系，这个 N 的变化形成了可调节大小的滑动窗口。最终将所有词语的概率累乘形成句子的概率，这样加强了词语之间的联系，但向量仍然稀疏，容易造成维度灾难。

TF-IDF 是一种关键词抽取方法，TF 是统计文本中的词频，IDF 为逆文档频率，词语在各文档中综合频次越高，IDF 值则越小，衡量词语的常见程度。相关计算公式如下：

（2） 分布式表示方法

词嵌入[2]是当下最流行的词向量表示方法。这种方法通过神经网络训练而来，在建模时考虑了上下文关系，生成的词向量有更丰富的语义信息。

[1] Word2Vec

Word2Vec 模型[4]是由 Mikolov 在 2013 年提出的，它包括两种训练方式：CBOW(Continuous Bag-of-words)和 Skip-Gram。CBOW 模型基本思想是从上下文中学习目标单词的分布式表示。具体来说它通过一个单词上下文窗口内的单词作为输入，预测中心单词。Skip-Gram 恰恰相反，尝试从中心单词学习上下文单词的分布式表示，二者原理如图 7。一般来说，Skip-Gram 在大规模语料库上通常表现更好，所以本项目采用 SGNS 模型。

SGNS 模型使用的是 Skip-Gram 模型，只是在训练过程中基于负采样（Negative Sampling）方法。在这个方法中，对于每个训练样本（一个中心词和它周围的上下文单词），会随机选择一些不相关单词作为负样本混合进正样本同时训练。

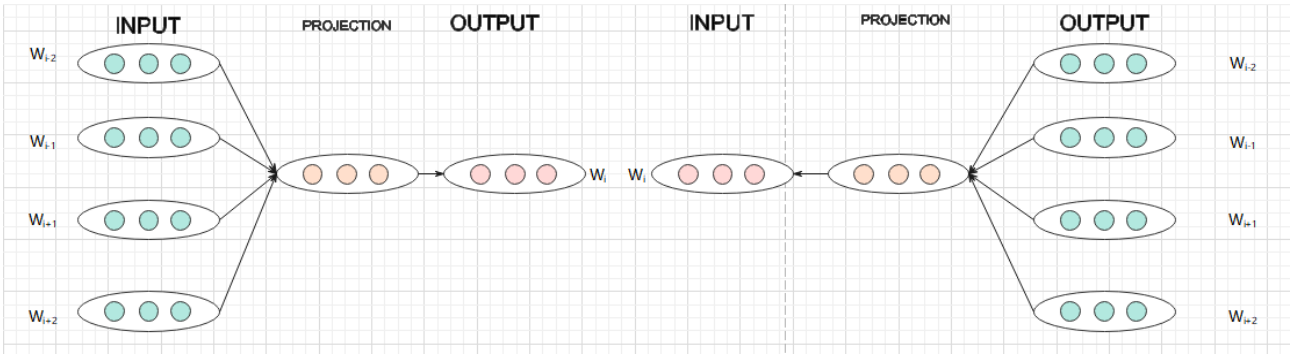

图 7 CBOW（左侧）模型与 Skip-gram（右侧）模型

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

图 8 是本项目“网络婚恋、交友类”[6],“虚假投资理财分布”两个类关键词向量化后的分布：

图 8 词向量可视化

我们可以看出两类的代表词向量明显分布为两簇，说明这两类的关键词在嵌入空间中分布有明显特征区别，所以这样的数据空间会让模型更好学到两个类别之间不同的特征表示。

[2] ELMO

ELMO 是 2018 年 AI2 提出的一种新的词嵌入方式，与传统词嵌入（Word2vec 和 Glove）不同，ELMO 能为每个单词生成多维度的嵌入向量，这些向量会根据单词在特定上下文中的语义和语法信息而变化，这解决了一词多义的问题。ELMO 通过大规模的语料库在双向 LSTM 或者双向 GRU 等双向语言模型上训练，这种模型通过编码、解码的方式学习到单词在不同上下文中的语义信息。

[3] BERT

同在 2018 年，Transformer 的变体（实际上是 Transformer 的 Encoder 层的纵深），预训练领域的里程碑之作 Bert 被提出，其相对 LSTM 这类模型具有了更好的特征提取能力。在预训练阶段，BERT 模型内部的嵌入层将输入文本的标记（单词、CLS、SEP 等）映射为动态嵌入向量。这些向量在 BERT 的各个层中会不断地被调整和更新，以便在多层次的语义和语境信息中寻找最佳的表示。

这些动态生成的嵌入向量是基于上下文的，因为它们同时受到输入文本的语境信息和 BERT 模型参数的影响。在预训练阶段，BERT 学会了如何将文本序列中的每个标记映射为上下文相关的动态嵌入。在预训练完成后，这些嵌入向量可以直接用于文本分类任务，也可以用在其他模型（如 TextCNN）中作为特征表示。

3.3 模型

将文本转化为能很好表示其语义的向量后，需要选择合适的模型学习这些向量中的特征，从而获得分辨文本[8]类别的能力。

3.3.1 TextCNN

短语的特征在文本分类中也很重要，CNN[7]作为作为局部寻优的模型，在捕捉这类特征有着很优越的能力。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。其模型结构大致分为输入层，计算层和输出层，如图 9。

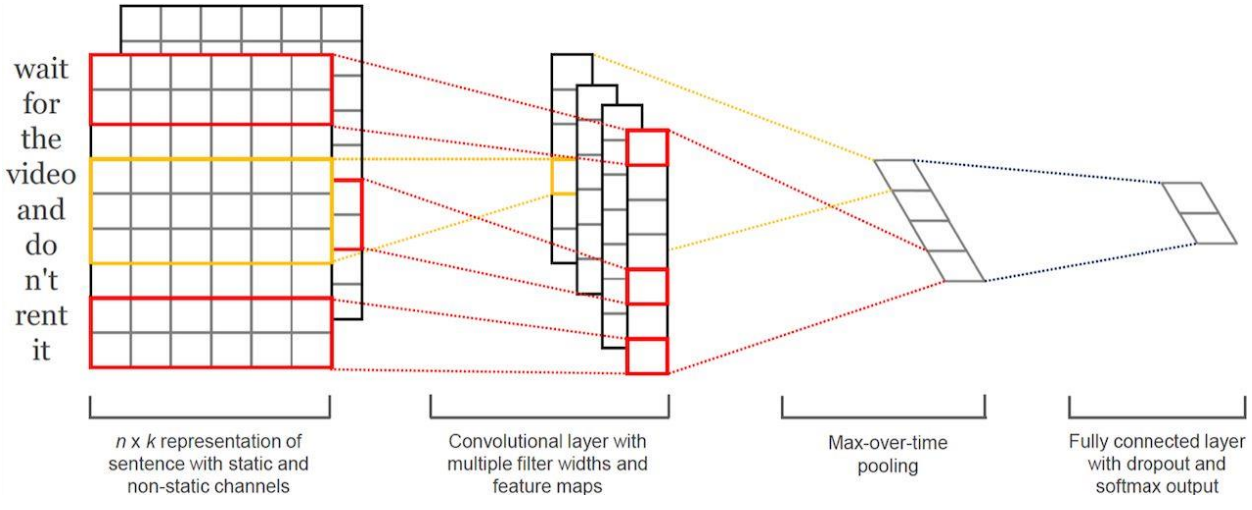

图 9 TextCNN 网络结构

输入层通过 Word2vec 词嵌入后，一个长度为 N 的句子被编码成 N×K 的二维特征矩阵，再通过 padding 操作，让所有句子长度都删补到 N。计算层由一个卷积层和一个池化层以及全连接层构成。本项目选取卷积核为（3，4，5），卷积计算后拼接三个不同维度的特征图，然后再通过池化层和全连接层。然后得到的 logits 再通过 softmax 得到最终的概率输出。

3.3.2 FastText

FastText 是 Facebook2016 年开源的一个词向量计算和文本分类工具，在文本分类任务中，FastText（浅层网络）往往能取得和深度网络相媲美的精度，却在训练时间上比深度网络快许多数量级。如图 10，FastText 模型也只有三层：输入层、隐含层、输出层。

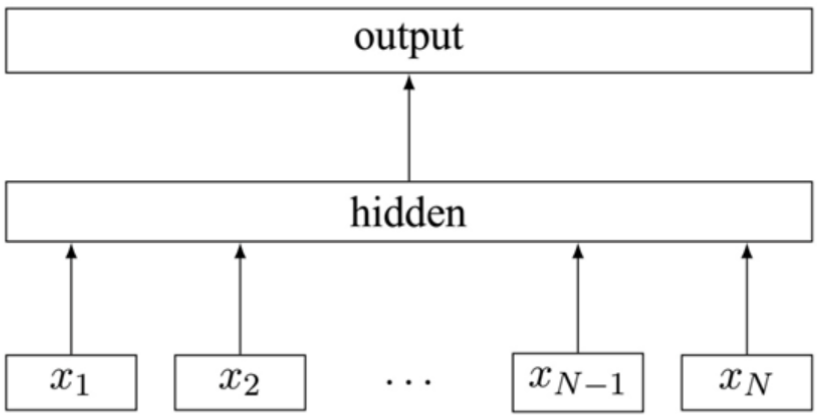

图 10 FastText 网络结构

输入都是多个向量化的单词，输出是一个特定的 target,隐藏层就是对多个词向量进行叠加平均。值得注意的是，FastText 在输入时将 n-gram 向量作为额外的特征；在输出时，

FastText 采用了分层 softmax,大大降低了模型训练时间。分层 Softmax 的基本思想是使用树的层级结构替代扁平化的标准 Softmax，使得在计算 P(y=j)时，只需计算一条路径上的所有节点的概率值，无需在意其它的节点。树的结构是根据类标的频数构造的霍夫曼树。K 个不同的类标组成所有的叶子节点，K-1 个内部节点作为内部参数，从根节点到某个叶子节点经过的节点和边形成一条路径，路径长度被表示为 L(y_j)。于是，P(y_j)就可以被写成：

$$p(y_j) = \prod_{l=1}^{L(y_j)-1} \sigma\left(\left[n(y_j, l+1) = LC\left(n(y_j, l)\right)\right]\right) * \theta_{n(y_j,l)}{}^{T} X \tag{2}$$

3.3.3 TextRNN

与 CNN 不同，RNN 模型天然具有捕捉文本时序信息[5]的能力，这种结构（如图 11）任意时刻的状态都与前 N 个状态相关，而这正适合用于长文本的特征获取。该模型的输入为$X = (x_1, x_2, \ldots, x_T)$，每个时间步的隐藏状态$h_t$ = RNN（x_t，h_{t-1}），以及输出序列$O = (o_1, o_2, \ldots, o_T)$。以下是每个单元的计算公式和模型整体结构图：

$$h_t = tan\, h(W_{hx} x_t + W_{hh} h_{t-1} + b_h) \tag{3}$$

$$o_t = W_o h_t + b_o \tag{4}$$

W_{ht}、W_{hh}、W_{Ot}为训练参数矩阵，b_t和b_{Ot}为偏置参数。tanh(*)为一种非线性激活函数。

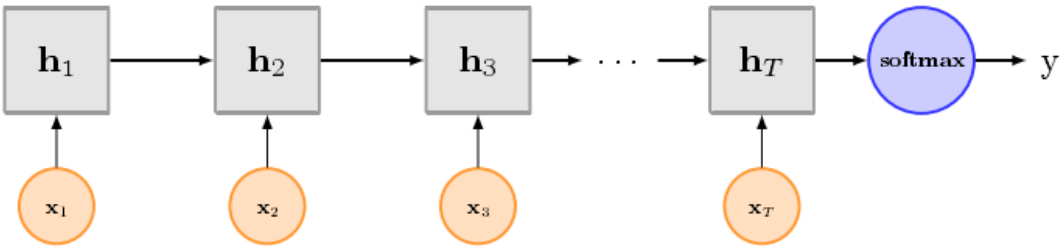

图 11 TextRNN 结构

由于信息量的不断增加，这样的结构会出现梯度消失或者梯度爆炸，所以本项目还采取了加 attention 的 RNN（实际实现的单元是采用 LSTM）, attention 的作用可以让模型注意到更需要注意的部分，如少样类等，防止过拟合增强模型的表达能力。

3.3.4 Bert

Bert 的输入一般包含了三个类别的嵌入，Token Embeddings、Segment Embeddings、Position Embeddings，其中 Position Embeddings 使用可训练的相对位置编码（Sinusoidal Encoding），很好的表示了文本的位置信息。输入再通过多个 Transfomer 的 Encode 模块实现文本特征的捕捉。从实践来看，这是很长一段时间的 SOTA 模型。以下是其公式说明:

对于一个输入序列$X = (x_1, x_2, \ldots, x_T))$，BERT 使用 Self-Attention 来计算每个位置 x_t 的上下文表示h_t：

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{5}$$

其中，$Q = XW_Q$，$K = XW_K$，$V = XW_V$是输入序列 X分别通过线性变换得到的查询（Query）、键（Key）和值（Value）矩阵，W_Q, W_K, W_V)是可学习的权重矩阵，d_k是查询和键的维度。

然后再使用多个自注意力头（h个）接受输入，每个头学习到不同的表示。这些头的输出被拼接，然后通过线性变换W_O得到多头注意力的输出：

$$MultiHead(X) = Concat(head_1, head_2, \ldots, head_h)W_O \tag{6}$$

其中，$head_i = Attention(XW_{Qi}, XW_{Ki}, XW_{Vi})$ 是第 i 个注意力头的输出，$(W_{Qi}, W_{Ki}, W_{Vi}, W_O)$是可学习的权重矩阵。多头注意力的输出通过一个前馈神经网络进行处理：

$$FFN\big(MultiHead(X)\big) = ReLU(MultiHead(X)W_1 + b_1)W_2 + b_2 \tag{7}$$

其中，(W_1, b_1, W_2, b_2)是可学习的权重矩阵和偏置，ReLU 表示激活函数。这一步允许模型进行非线性变换。

BERT 通过以上步骤将输入序列转换为上下文相关的表示，这些表示可以用于各种自然语言处理任务。BERT 的预训练阶段采用了 Masked Language Model（MLM）任务和下一句预测任务，使得模型能够学习到丰富的语言表示。如图 12，在我们的任务中，这些预训练的表示可以被微调，以适应我们分类任务的需求。

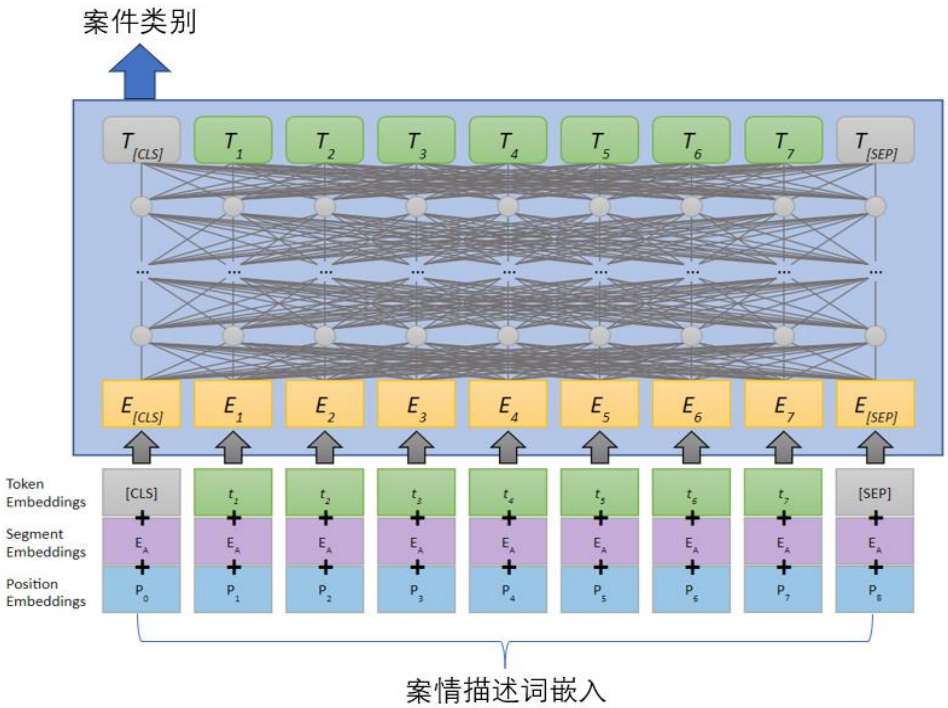

图 12 bert 结构

3.3.5 ERNIE

ERNIE 在 Bert 基础上，引入了世界知识（world knowledge）的概念，通过结合来自百科知识库等外部数据源的信息，提供了更丰富的语义表示。这种融合使得模型更好地理解文本中的实体、事件和关系。ERNIE 引入的知识融合层，用于将外部知识与文本语义融合。这个层可以包括多种不同的知识表示，例如实体、属性和关系。在 ERNIE 的训练过程中，知识融合层会自动学习如何利用外部知识。事实证明，其在我们的文本分类任务上表现得更好。

3.4 数据不均衡

我们根据一次的训练日志得到的测试集混淆矩阵，观察到几个现象：

- 冒充电商物流客服类有很多被分类到了虚假征信类
- 冒充军警购物类诈骗基本被分类到了虚假购物、服务类
- 虚假购物、服务类与多个类别有重叠
- 网络婚恋交友与多个类别重叠

基于此我们首先采取了降采样和过采样的方法，但通过简单的采样后发现重采样可能导致类内样本不均衡，这是因为不同类别中的样本数目可能本身就不同，而重采样方法通常是基于类别来进行采样的，所以它会在不同类别之间引入概率偏差，导致类内样本的不平衡。这对我们进行重采样的精度有很大的要求，然而带来的提升并不显著。

最终，我们提出解决方案—合并标签：

（1）可以通过减少类别将虚假征信类与冒充电商物流客服类合并成冒充客服类，数据集中虚假征信类的数据中绝大部分都是京东金融、白条、支付宝等客服进行诈骗

（2）虚假购物服务类与网络游戏产品虚假交易类以及冒充军警购物类诈骗两个有一定的重合，而网络游戏产品虚假交易类和冒充军警购物类数量较少，均与购物有关，所以可以合成一个虚假购物服务类

（3）网黑案件类中基本上都是网上认识人然后裸聊的案件，也可以与网络婚恋、交友类合到一起

3.5 模型校准

考虑到对于数据重叠的问题，也可以考虑成一条数据有多个标签，此时我们可以将 logits softmax 之后的值排序输出，但是通过观察，发现 TextCNN 对于自己的判断 过于自信（over-confidence）。即使预测错误，logits 中最大类的概率值和其他类差距过大，这表明模型不知道自己预测错误。所以我们考虑对模型进行校准，得到一个可靠的 logits 来作为模型的 置信度（confidence）。

3.6 模型融合

在前端的预测中，我们还提供了集成后的模型 ensemble, 该模型通过不相同数据集训练出的 k 个模型（即本项目中的 TextCNN，FastText,TextRNN,bert,ERNIE）分别预测结果，然后通过简单的 logits 相加，得到最终的 logits 选出概率值最大的 label，从而得到更合理的结果。经过实验，模型性能有一定提升，模型可靠性有显著提升。

3.7 实验与评价

3.7.1 评价标准

本实验的评价标准分为两方面：性能评价和模型置信度评价。在性能评价方面，我们采用测试集的准确率（accuracy，简称 acc）作为评价标准，它衡量的是模型预测正确的样本数占总样本数的比例。在模型置信度评价方面，我们使用了期望校准误差（Expected Calibration Error，简称 ECE）进行评估。ECE 是一种衡量模型预测概率准确性的指标，它计算的是模型预测的概率和实际发生的频率之间的差距。通过这两个指标，我们可以全面地评估模型的性能和置信度，从而更好地理解和优化我们的模型。

其中 ECE 指标的计算公式为：

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{n} |acc(B_m) - conf(B_m)| \tag{8}$$

其中，M 是分箱的数量，Bm 是第 m 个箱子中的样本集合，｜Bm｜ 是第 m 个箱子中的样本数量，n 是总样本数。acc(Bm) 是第 m 个箱子中样本的准确率， conf(Bm) 是第 m 个箱子中样本的平均预测概率。ECE 是用于评估分类模型的校准性的指标，数值越低表示模型的预测越准确。

3.7.2 评测结果

表 3 各个模型评测结果

模型	12 分类 acc	8 分类 acc	12 分类 ECE	8 分类 ECE
TextCNN	86.82%	91.70%	0.6790	0.6420
FastText	87.18%	90.40%	0.6843	0.6463
TextRNN	86.39%	89.85%	0.6776	0.6388
Bert	86.40%	90.36%	0.6909	0.6488
ERNIE	88.40%	91.89%	0.5797	0.5214
Ensemble	88.58%	91.92%	0.0705	0.0385
GPT_FT	\	92.01%	\	\

模型性能比较：

- 在 12 分类准确率方面，Ensemble 模型表现最好，达到了 88.58%的准确率，相较于其他单独的模型更具优势。
- 在 8 分类准确率方面，同样是 Ensemble 模型表现最出色，达到了 91.82%的准确率。

- ERNIE 模型在 12 分类和 8 分类准确率上都表现得非常好,分别为 88.40%和 91.89%。
- TextCNN、FastText、TextRNN 和 Bert 在准确率上表现相对较为一致，但都比 Ensemble 和 ERNIE 略逊一筹。

ECE 分析：

- 在 12 分类 ECE 方面，Ensemble 模型表现得相当出色，为 0.0705，远远低于其他模型。
- 在 8 分类 ECE 方面，同样是 Ensemble 模型的 ECE 值最低，为 0.0385，意味着其在 8 分类问题上的预测非常接近真实标签。
- 所以模型集成是解决模型过度自信的一个很好的办法

总体分析：

- ERNIE 在综合 12 分类、8 分类准确率和 ECE 指标上表现得非常出色,说明 ERNIE 是一个很出色的模型。
- TextCNN、TextRNN、Bert 在准确率上表现稳定，但在 ECE 指标上相对较高，说明其预测的概率分布与实际分布有些偏差。
- 相较于其他单一模型， Ensemble 在性能上和可靠性上都占据绝对优势，说明模型集成带来了性能巨大飞跃。
- 但是由于本项目训练资源有限，实际训练 Bert 和 ERNIE 时参数 batchsize=32, padding=128,训练可能不充分，没有完全发挥这两个预训练模型的能力，所以这只能是在低资源下的分析结论。
- 此外本项目还使用 gpt 官方 api 微调 gpt 基本模型进行文本分类，只使用了 2000 数据且选择最最便宜和最快的 ada 模型,便超过了本项目 8 万数据训练出的 SOTA，可以看出，这样海量参数的通用大语言模型对于解决我们这种问题是十分轻松的，未来我们也会在这方面更多探索。

3.8 可视化分析

3.8.1 Flask 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，部署上服务器。Flask 框架，轻量级，扩展插件较多，且其有超高的扩展性和小而精的核心，所以项目将使用 Flask 框架。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。工作过程如图 12 所示：

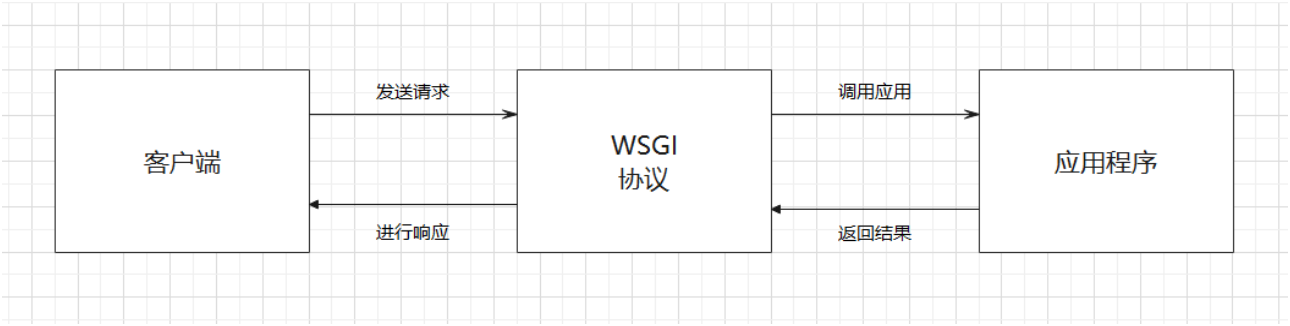

图 13 Flask 工作过程图示

3.8.2 MySQL 数据库

MySQL 是目前最受欢迎的开放源码关系数据库系统。MySQL 具有性能好、易学、开源等优势，被各类企业以及独立开发人员所使用，同时其速度快、体积小、成本低。基于上述原因，在进行此项目的设计和实现时，使用 MySQL 用于存储和管理数据。

数据库设计：诈骗案件实体包括了类别编号、用户反馈、案情描述、案件类别属性，如图 13 所示

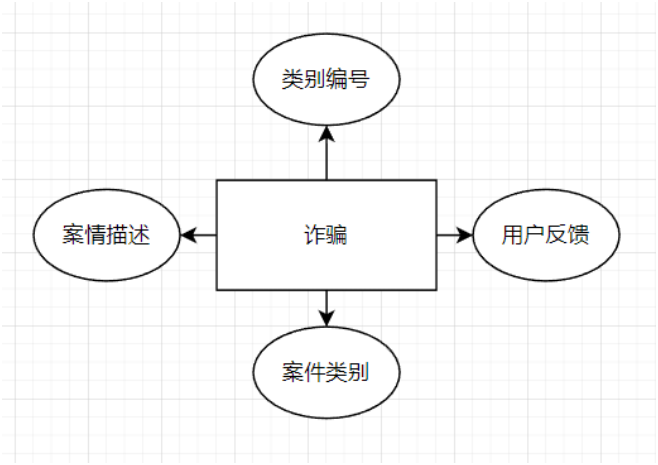

图 14 诈骗案件实体图示

数据库格式

表 4 数据库逻辑设计

字段	注释	类型	非空
completion	类别编号	int	是
prompt	案情描述	varchar(10000)	是
classes	案件类别	varchar(32)	是

like	用户反馈	varchar(10)	是

3.8.3 前端设计

我们已经编写了三个 HTML 网页，分别是核心文本分类项目的主页（index.html）、基本教程展示页（start.html），以及有关项目图表的展示页（chart.html）。

1. index.html 主页

主页实现了预测的基本功能，用户可以通过下拉框选择要使用的文本分类模型。以下是主页的关键设计和功能：

- 输入框将姓名、身份证以及户籍与案情描述进行了分离，仅将用户输入的案情描述传回后端，以减少无关信息对预测结果的可能影响。
- 页面下方包含三个模块：Quick Start（简单教程）、Reference Project（参考项目）、Charts（图表），以提供更全面的信息。
- 左上角悬浮的电子时钟方便用户查看当前时间和记录备案时间。
- 文本框和 Predict 按钮添加了点击动画，提高网站的观赏性。

用户在输入案情描述后，通过下拉框选择一个模型，提供了八个选项：FastText、TextCNN、TextRNN、BERT、ERNIE、ChatGPT、GPT_FT、Ensemble。其中，Ensemble 通过前五个模型进行预测，根据它们的结果选择出现次数最多的一个作为集成预测结果。最近由于 ChatGPT 的广泛应用，我们接入了 ChatGPT 接口，同时还尝试使用 GPT api 和我们的数据微调一个分类模型 GPT_FT, 以提供更准确的分类结果和相关建议。

用户选择好模型后，点击预测会在下方输出对应的预测结果。页面的下半部分显示了对应诈骗类型的分析、反诈警醒，以及案情描述中的一些关键词，提醒读者警惕敏感词汇。

在预测结果下方设置了两个按钮，用于用户反馈结果的正确性。如果用户点击“点赞”按钮，该结果和案情描述将直接存储到数据库中，并标记为“yes”。点击“踩”按钮会弹出一个窗口，用户可以选择认为正确的类别，点击确定后，将该案情描述和用户选择的案件类别存储到数据库中，并标记为“no”。这使得系统通过不断地从用户反馈中学习，逐步提高准确性。

我们会定期利用这些新数据重新训练系统，以确保它与最新的欺诈趋势和模式保持同步。

图 14 左侧为预测前的 index.html 主页，右侧是预测后的 index.html 主页。图 15 是用

户点击踩之后的弹窗。

图 15 index.html 页面图示

图 16 点击“踩”按钮后弹窗图示

2. chart.html 页面

在 chart.html 页面中，我们进行了对项目结果的可视化分析，以下是对图表和可视化内容的详细描述：

[1] 词云图

- 图 16： 词云展示了文本中出现频率最高的 120 个词汇。这有助于读者了解文本中的主题和关键词，特别是对于敏感词汇的预警。当读者碰到这些词汇时，应该引起警惕。

[2] 案件类型分布图

- 图 17： 扇形图展示了案件类型的分布情况，通过直观的方式呈现了各个类别的占比。其中刷单返利类占据最大的比例。这有助于读者快速了解各类别之间的相对频率。

- 图 18： 柱状图也展示了案件类型的分布情况，提供了更加详细的分类数据。这有助于更深入地了解案件类型的相对比例，使得数据更有价值。

[3] TensorBoard 可视化

- 利用 TensorBoard 实现了准确率和损失的动态可视化，包括了训练准确率与损失、验证准确率与损失。这样的可视化能够帮助用户直观地了解模型的训练情况，以及在不同阶段的性能表现。
- 为了更详细地了解每个模型的性能，加入了各个模型准确率的仪表盘。这提供了一个快速参考，帮助用户选择适合其需求的模型。

这些可视化工具不仅使得数据更容易理解，而且提供了对项目整体性能和关键指标的直观把握。用户能够通过这些图表更深入地了解文本分类的结果，进而做出更明智的决策。

图 17 文本中出现频率最高的 120 个词

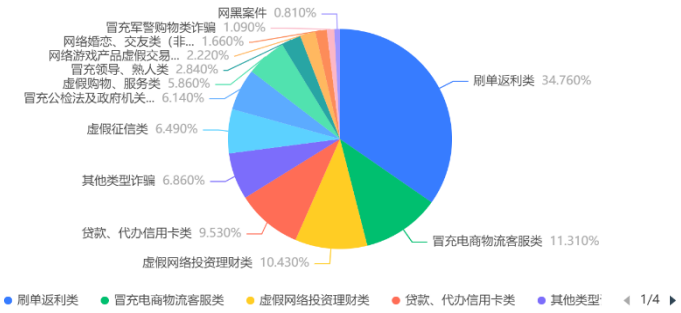

图 18 案件类型分布情况扇形图

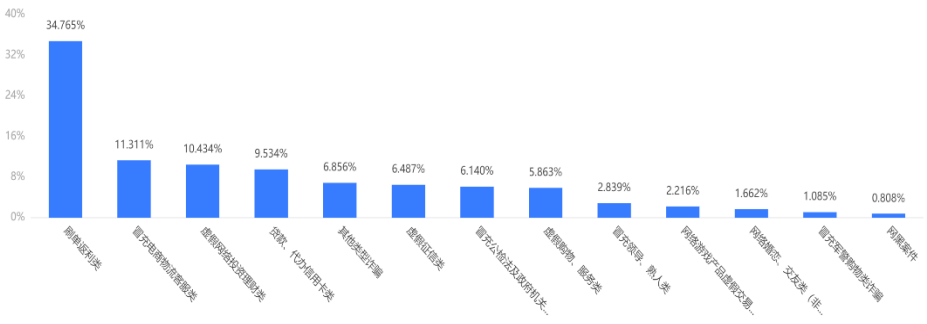

图 19 案件类型分布情况柱状图

我们通过购买腾讯云的轻量应用服务器，服务器规格为 CPU 2 核、内存 4GB、系统盘 60GB。这个配置对于我们的项目来说提供了足够的计算和存储资源。将本项目部署到该服

务器上，便于用户通过网络直接访问我们的项目提高了可用性和可访问性，这对于用户远程访问和使用项目是非常重要的。也便于我们进行管理与维护。此外云服务器环境提供了更高水平的安全性，我们按照最佳实践配置了防火墙、安全组等安全措施，通过腾讯云提供的安全设置和防护机制来保护项目免受潜在的攻击和威胁。同时也提高了我们项目的可扩展性，可以通过增加服务器的实例、负载均衡等方式来提高系统的扩展性。

4、结论（成果介绍）

本项目基于电信网络诈骗案件分类任务，开发了一个精度高，易使用，能实用的诈骗文本分类系统。此系统包含数据库，多模型的模型层，以及多功能易于使用的前端网页。通过一系列的优化，包括数据的分析和处理，类别融合，多模型集成，以及模型校准，取得了显著的成果。

通过细致的数据分析和优化和类别融合，模型提取了更精确的特征，有助于准确地检测电信网络诈骗案件。同时，通过多模型集成，增强了模型的鲁棒性，提高了分类的准确性。同时基于 flask 框架和 MySQL 数据库，构建了一个集成各模型，包含各类别分析的前端交互网页。

5、经费使用情况

表 5 经费支出表

预算类别	主要用途	预算金额（元）
书籍费	购买书籍等	300
资料打印	打印所需的资料	50
服务器租聘	模型训练，网站搭建	400

6、问题、体会与收获

在这个电信网络诈骗案件分类项目中,我们有了很多思考。在数据处理方面, 我们意识到了数据不平衡问题的严重性,不同类别之间和内部样本分布不均是我们需要重点关注的问题。我们学习了重采样、减类等方法来处理这一问题。在特征提取方面,我对各种文本表示方法如 BOW、词向量、BERT 等有了直接的体验和比较,知道了需要根据具体任务选择合适的方法。在模型选择上, 我们发现集成多个不同的模型可以很好地提升分类的鲁棒性和准确率。另外,仅看准确率是不够的,我们还需要关注模型预测概率的可靠性,这需要通过校准来提高。最重要的是, 我们体会到了团队协作的高效性,通过团队的协作和合作,我们能够更好地完成项目,提高效率,确保项目的质量,实现项目的目标。通过这个项目, 我们也对

nlp 各种任务有了较为系统的理解,也对实际任务处理有了进一步的认识。我们收获了宝贵的文本表示、模型构建、结果评估等方面的实践经验。这些都将帮助我更好地开展今后的科研工作。

7、建议

针对这个电信网络诈骗案件分类项目,我们觉得这个项目还有很多方面可以进行扩展和优化:第一,可以尝试采用更先进的图神经网络等特征学习方法,以建模文本的语义关系,提高分类性能。第二,可以引入传统机器学习方法如 SVM 进行特征工程,与深度学习模型形成 Ensemble,增加鲁棒性。第三,收集更多样例数据,优化样本分布,解决数据不平衡问题。第四,利用相关领域的数据进行迁移学习,迁移源领域的知识来增强目标任务。第五,加入规则方法来补充深度学习的不足。第六,从人机交互角度优化前端界面,提高系统的智能化和友好性。第七,实际部署服务,收集用户反馈进行持续改进。第八, 机器学习在反欺诈领域无疑可以发挥重要作用,但同时也可能被不法分子利用,所以从攻防的角度来研究这个问题也十分有必要。

8、结束语与致谢

经过一年左右的不懈努力，我们的项目基本实现了申报大创的目标。未来，我们将持续关注类似话题，通过不断的研究与实践，进一步提升系统的实用性和性能。我们坚信，这项项目的成果将为社会带来实实在在的利益，提高诈骗案件的识别能力，保护公众的合法权益。

最后，我要特别感谢我们的指导老师孙承杰老师。在项目进行的过程中，我多次向孙老师请教问题，也多次在学习生涯的规划上向他请教，感激他每一次耐心回答和悉心指导。同时，我也要感谢指导本项目的纪杰学长。在项目和学习上的困惑，学长用同龄人的口吻帮助了我们许多，有了学长的帮助，我们进步飞速。在此，我向老师和学长表示最诚挚的谢意！

此外，我要特别感谢与我一起完成项目的团队成员。正是我们的坚持与努力，使得项目最终取得了圆满的结果。希望我们在未来的学习和科研生涯中继续共同前行，互相激励，共同成长。

9、参考文献

[1] 陆家炜, 田翀, 李博文, 沈久一, 胡佳伟, 关琳. 公安案件电子卷宗文本计量研究——以电信诈骗案件为例[J]. 数据挖掘, 2020, 10(3): 221-228. https://doi.org/10.12677/HJDM.2020.103023

[2] Shen Li, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, Xiaoyong Du, *Analogical Reasoning on Chinese Morphological and Semantic Relations*, ACL 2018.

[3] 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.

[4]Tomas Mikolov,Kai Chen 0010,Greg Corrado,Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space[J]. CoRR,2013,abs/1301.3781.

[5]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.

[6]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.

[7]Yoon Kim. Convolutional Neural Networks for Sentence Classification.[J]. CoRR, 2014,abs/1408.5882.

[8]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.

六、附件（专利、发表论文及其他成果支撑材料）