多模态融合及特定损失函数设计在涉诈网站中的应用

摘要

本文介绍了多模态融合技术和特定损失函数设计在涉诈网站中的应用。这项研究对于当前社会而言具有重要的现实意义，涉诈网站已然成为了网络安全领域的一大难题，造成了极其严重的社会影响和经济损失。

将多模态融合与特定损失函数应用于涉诈网站的检测中，可以提高模型的性能，准确地识别诈骗网站，避免更多的人受到骗局的影响。这个项目的意义在于，通过使用机器学习算法实现对涉诈网站的自动化检测和预警，能够提高防范涉诈活动的效率和准确性，并为打击网络诈骗提供有力的技术保障。

为了满足有层次的服务选择，我们对于不同需求的网址分类提供了不同数量级的数据支撑。我们主要提取了词法特征，主机特征和内容特征三个维度的特征，为模型构建数据集。

为了解决现有建模中存在的三个缺点，我们提出了使用特定损失函数正则化数据分布特征，并对视觉-文本多模态信息以及网站主机特征、网址 ULR 特征进行融合。

同时，为满足特定场景下对模型推理速度的要求，提供了两种模式：正常模式和高质量模式。关于损失函数设计方面，文章中采用了 Local loss 对数据集中占比过大的正常网址数据进行处理，以提高涉嫌诈骗网址的回召率，保证模型的预测能力。

在文本部分的处理上，我们采用了爬虫获取网站文本数据，并对 BERT-wwm 模型在该模式下进行二次预训练。考虑到文本长度通常会超过 BERT-wwm 输入文本的最大长度，为了解决这个问题，作者采用了直接截断的方法对文本进行处理。

此外，对于图片部分，我们采用 ResNet 对爬取的网页快照进行卷积，并将卷积获得隐藏层输出作为图片部分的特征，进一步提高了模型的性能和准确率。

最终，文章提出的多模态融合及特定损失函数设计在涉诈网站中的应用的模型在推理阶段的架构如文章所示，实现了涉诈网站的自动化检测和预警，具有非常重要的实际应用价值。

目录

1 项目背景

随着互联网技术的飞速发展，人们的日常生活已经离不开互联网。网络让信息传递更加迅速、方便，为人们提供了丰富的资源。然而，随着互联网的普及，网络安全问题也日益突出。网络犯罪手段层出不穷，给用户的信息安全和网络环境带来了严重威胁。

恶意网址是指那些用于传播恶意软件、钓鱼攻击、欺诈活动等恶意行为的网站。这些网站通常会模仿正规网站，以诱导用户访问。恶意网址的存在严重破坏了网络安全，导致用户财产损失、个人隐私泄露等问题。因此，研究如何有效识别和分类恶意网址，提高用户的网络安全意识和防范能力，成为亟待解决的问题。

尽管目前已有一些安全软件和浏览器插件可以帮助用户识别恶意网址，但仍存在以下问题：(1) 恶意网址识别率不高，导致一些恶意网址仍能逃过检测；(2) 部分恶意网址因为页面内容变化而逃避识别；(3) 针对恶意网址的应对措施不足，用户在面对恶意网址时缺乏有效防范方法。

为了解决这些问题，本项目旨在设计一套高效、可靠的恶意网址识别与分类算法，提高网络安全水平，帮助用户及时发现和防范恶意网址，降低网络犯罪的发生率，保护用户的信息安全。此外，通过对恶意网址的分析，我们可以了解其发展趋势和特点，为网络安全领域的政策制定和技术研发提供有益参考。

2 国内外研究现状

目前，有两种主要的方法用于自动涉诈网址分类模型的研究。第一种是基于特征工程 + 机器学习分类模型的方式，主要是通过设计良好的特征来实现对恶意网址的自动分类。第二种是使用深度学习算法，如卷积神经网络 (CNN) 和循环神经网络 (RNN)，来进行恶意网址的自动分类。国内外在这两种方法上都有相关研究。

2.1 国内研究现状：

贵州大学某团队借助文本分类算法对涉及信用卡的诈骗网站进行识别

清华大学朱建伟团队使用深度学习方法来判定网址是否恶意

2.2 国外研究现状：

Alazab, et al. 提出了一种新的模型，称为 “URL Ranker”，它使用半监督机器学习分类器来提高其分类效果，该算法以最小化错误分类为目标，同时最大化对未知 URL 的检测能力

Pascanu, et al. 使用长短时记忆网络 (Long Short-Term Memory, LSTM) 神经网络对恶意 URL 进行分类，LSTM 可以自适应地学习有关 URL 内容和上下文信息的复杂模式，并获得更好的性能

Zhang, et al. 借助 LightGBM 模型对恶意 URL 进行了分类，该模型通过某些特征向量，例如 URL 长度、网页大纲、数字和字符等特征，以及进行四个月以上学习的白名单和黑名单词典，来判断给定的 URL 是否是恶意的

3 项目需求分析

3.1 用户需求

- 用户希望能够在访问网站时，算法能在访问网站前判断网址类型以及网站是否恶意，防止受到网络攻击。
- 算法需要具有高准确率的恶意网址识别能力，避免将正常网站误判为恶意网址，同时确保真正的恶意网址被有效识别。
- 算法能够判断恶意网站的具体类型，了解恶意网址的具体类型（如钓鱼、恶意软件分发等），以便采取相应的防范措施。

3.2 技术需求

- 数据收集与处理：技术需求包括从多种来源收集大量的网址数据，进行预处理和标注，确保数据质量和可靠性。
- 特征工程：从网址的词法特征、内容特征、主机特征等多个维度提取有效特征，为模型训练提供有力支持。
- 模型设计与训练：采用先进的机器学习和深度学习方法，构建恶意网址识别与分类模型，并进行训练和优化，提高模型的准确率和泛化能力。
- 模型评估与优化：持续关注模型在的表现，根据真实数据和需求进行优化，提高算法性能。

3.3 市场需求

- 市场潜力：由于网络安全问题日益严重，打击恶意网址识别与分类项目具有广泛的市场需求和潜力。
- 竞争分析：了解市场上现有的恶意网址识别与分类产品，分析其优势与不足，以指导本项目的发展方向。
- 法规遵循：遵循相关法律法规，确保项目的合规性，避免因侵犯用户隐私等问题导致的法律风险。

综合以上需求分析，打击恶意网址识别与分类项目需要在用户需求、技术需求和市场需求三个方面进行深入研究。项目的核心目标是设计一套高准确率、易用性强的恶意网址识别与分类算法，以提高网络安全水平，保护用户的信息安全。在技术实现方面，项目需关注数据收集与处理、特征工程、模型设计与训练、系统部署与集成等环节。此外，项目还应关注市场需求，挖掘潜在的市场机会，确保项目在实际应用中发挥最大价值。

4 建模思路

4.1 现有模型存在不足

- 没有考虑到数据集分布对模型训练的影响
- 没有考虑到使用多模态特征分类准确率
- 没有考虑到使用可伸缩的网站预测模式

在本文中的建模中，我们考虑到以上三个缺点，提出了使用特定损失函数正则化数据分布特征，同时对视觉-文本多模态信息以及网站主机特征，网址 ULR 特征进行了融合，并根据特定场景下关于模型推理速度的要求提供了两种模式：正常模式和高质量模式。

4.2 损失函数

在损失函数设计方面我们采用了如下损失函数：

$$\sum_{x\in\mathcal{D}} \mathbb{I}_{y\in\mathcal{Y}^+}\rho \log p(y \mid \theta, x) + \mathbb{I}_{y\in\mathcal{Y}^-} \log p(y \mid \theta, x) \tag{1}$$

采用此方法的原因是数据集中存在占比过大的正常网址数据，而我们的目的是提高涉嫌诈骗网址的回召率，因此需要类似于 Local loss 那样提高负样本在训练过程中的权重。

4.3 文本处理

关于文本部分的处理我们采用爬虫爬取网站的文本数据，并对 BERT-wwm 模型在该模式下进行二次预训练。在训练下游的分类任务的过程中，由于文本长度通常会超过 BERT-wwm 输入文本的最大长度，所以需要对文本进行摘要处理。一种方法是采用直接截断的方法处理，另一种方法是采用例如 TEXTRANK 的方法提取摘要。这里我们比较了两种方法的差异，发现在第一种情况下通常就可以包含文本的主题，出于对推理速度的考量我们使用了直接截断的方法。

4.4 图像部分的处理

关于图片部分的处理我们采用 ResNet 对爬取的网页快照进行卷积，并将卷积获得隐藏层输出作为图片部分的特征

4.5 其它主机特征

其他关于主机特征和网址 URL 的特征可以参考数据处理部分

4.6 系统架构

最终我们的模型在推理阶段的架构如图1所示：

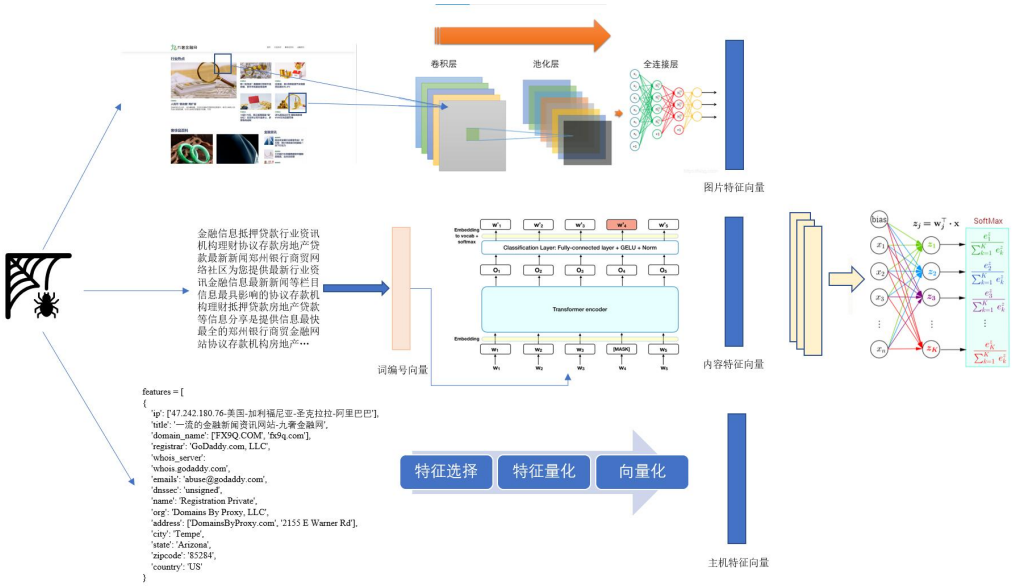

图 1: The Architecture of the Inference Phase

5 项目实施方案

5.1 数据收集

通过爬虫技术、网络安全机构提供的数据等多种途径，收集大量的网址信息，包括正常网址和恶意网址。

为了进行有效的涉嫌诈骗的网址自动分类识别，获取足够的特征是非常重要的，在大量数据的基础上才可以获得更好的模型。

为此，我需要获取多个维度的特征，主要包括词法特征，主机特征，内容特征，具体数据获取及处理部分如下：

第一个维度是词法特征，主要由 URL 字符串中的元素组成。它们是由 URL 的文本属性决定的，包括统计属性，如 URL 的长度，域名的长度，特殊字符的数量，以及 URL 中的数字数量；也包括引申属性，如域名的等级，域名中各个组件代表的含义等，同时也包括内部的隐含属性，如文本的复杂程度，字符之间的交互信息等。

第二个维度是主机特征，需要通过 url 请求获取特征，主要包括 ip 地址，标题，域名，登记，whois 服务，更新时间，创建时间，到期时间，status，邮箱，DNS 解析，组织，地点，国家，城市和邮编等信息。在这一部分中，我们主要通过 SearchMap 进行相关信息的获取，searchmap 是一款集域名解析、IP 反查域名、WHOIS 查询、CDN 检测、端口扫描、目录扫描、子域名挖掘为一体的前渗透测试综合信息收集工具，可以通过这个工具获取到上述的信息。

第三个维度是网页内容及其相关特征，主要包括文本内容和图片快照。

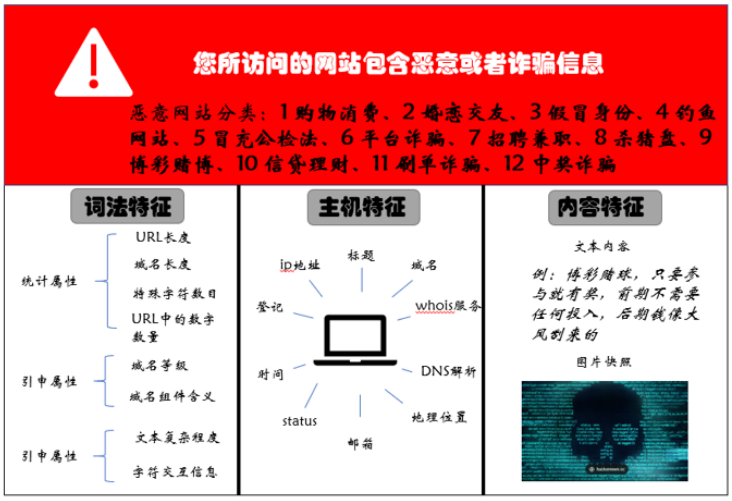

图 2: 数据特征维度展现

5.2 数据预处理

对收集到的网址数据进行清洗、去重、格式化等预处理操作，确保数据质量。

下面陈述数据处理部分

第一个维度是词法特征。数据处理时需要对于 url 本身进行处理，首先对于 URL 中的字符进行统计，获取所有的统计属性，这些统计属性都是连续性变量，因此无需进一步处理；然后需要获取 URL 的引申属性，由于域名是以英文符号 “.” 来隔开的，顶级域名是固定的，形如.com、.net、.org、.gov、.edu、.cn、.cc、.to、.tv 等，形如 http://baidu.com 为一级域名，形如 http://www.baidu.com 为二级域名，通过域名等级进行域名的划分，对于每一个域名的组成，通过分类指标进行转化；进一步需要获取对应的隐含属性，包括计算文本的复杂程度，字符之间的交互信息，主要通过 CNN 进行特征提取，考虑到 URL 中字符的顺序对于网址识别也具有重要意义，因此加一个 BiLSTM 层可以提取到顺序信息，通过最终的损失函数进行特征提取参数的优化，实现对于文本隐含信息的计量。

第二个维度是主机特征，数据的处理时需要分数值型特征和变量型特征分别进行处理，数值性特征主要包括三个时间，我们通过计算 2023 年 1 月 1 日 00：00：00 与所有的时间的差值，将时间转化为更好处理的数值型变量，对于域名，以 128.1.2.0 为例，我们将其以 ‘.’ 作为分割，转换为四个数值型变量，对于除了标题的其他信息，均看作分类变量进行处理。

第三个维度是网页内容及其相关特征，主要包括文本内容和图片快照。

5.3 特征工程

从网址的结构、内容、访问行为等多个维度提取特征，为模型训练和识别提供有效信息。

在涉嫌诈骗的网址自动分类识别项目中，获取合适的特征是非常关键的。为此，我们采取了大量的特征提取方法，包括以下几个方面:

i. url 特征：这些特征主要从 url 的结构和内容等方面提取信息，包括长度、字符编码、域名、子域名、路径、查询字符串、后缀等。这些特征可以帮助我们分析 url 的真实意图和目的，以进行分类。

ii. 网页截图特征：这些特征主要从网页的外观和页面元素等方面提取信息，如页面布局、颜色、字体、图片等等。这些特征可以直观地反映出网页的特点和风格，从而辅助我们进行分类。

iii. 网页文本特征：这些特征主要从网页文本内容方面提取信息，如标题、描述、关键词、正文内容等等。这些特征可以帮助我们在语义上理解网页的真实含义和用途，进而进行分类。

iv. 报文响应头特征：这些特征主要从 HTTP 协议通信报文中提取信息，如状态码、响应时间、服务器类型、Cookie 等等。这些特征可以帮助我们判断网站是否正常运行，从而进行分类。

v. whois 和 ALEXA 特征：这些特征主要从域名注册信息和流量统计等方面提取信息，如域名注册时间、所有者信息、预估流量、排名等等。这些特征可以辅助我们对网站进行更全面的分析和判断。

对于获取的数据，我们要进行合适的特征工程处理，将数据量化成模型可以处理的向量，首先，对于图片快照，最好的特征提取方法就是对于像素点的信息进行提取，然后通过 CNN 层进行特征的提取和聚集，具体示意图如下：

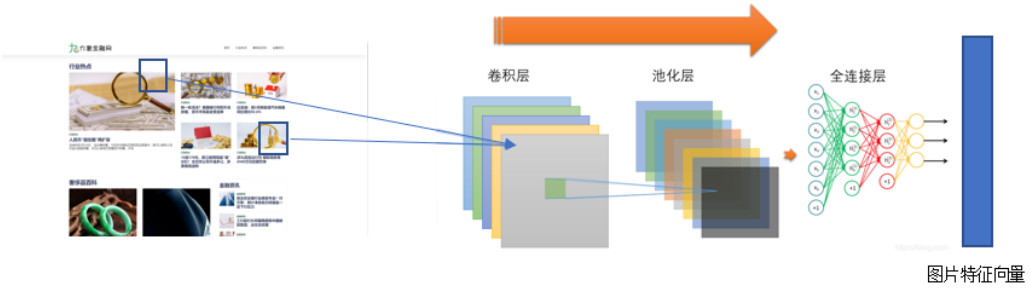

图 3: 图片向量构造过程

其次，对于文本信息，提取文本中的关键信息，对于网址的合理分类具有重要的意义，因此我们通过改进过的 Effentient Transformer 对于文本中的关键信息进行提取和量化，具体示意图如下：

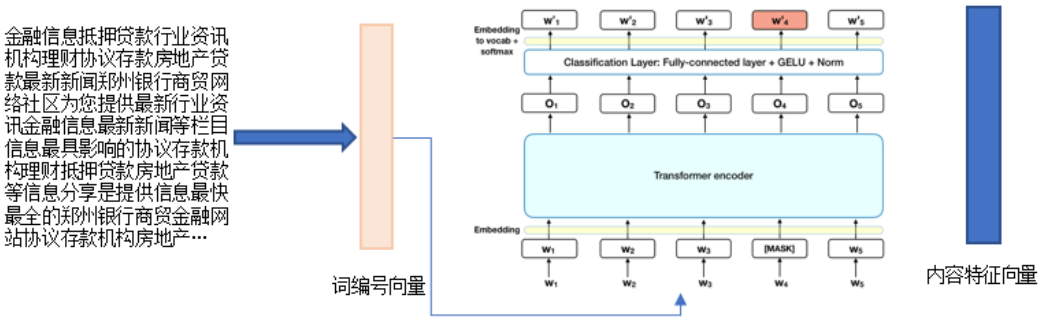

图 4: 内容向量构造过程

接着，我们将所有的特征向量进行向量拼接，对于模型特征再进行一定的特征筛选，得到可以供模型训练的完全特征向量，如下图：

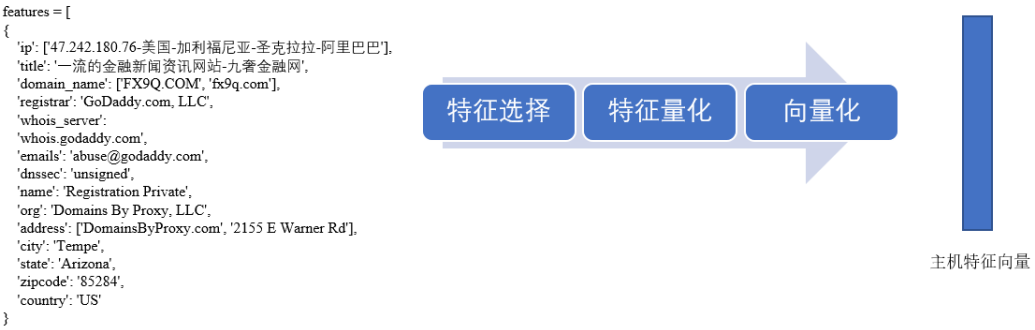

图 5: 特征向量生成过程

这些特征各自具有不同的重要性和区别，但在整个模型中都起到了至关重要的作用。我们综合应用这些特征，通过机器学习算法进行训练和优化，最终实现高效准确的涉嫌诈骗网址自动分类识别，至此特征工程完成。

5.4 模型训练与优化

采用机器学习和深度学习等方法，构建恶意网址识别与分类模型，并进行训练和优化，提高模型的准确率和泛化能力。

5.4.1 实验设置

CPU：Intel(R) Xeon(R) Gold 6226R CPU @ 2.90GHZ，32 线程，GPU：NVIDIA RTX A6000。

5.4.2 BERT 二次预训练

收集 Train.csv 中的部分网站地址的文本内容生成语料库进行二次预训练，预训练采用和 Bert-wwm 中相同的全词 MASK 任务进行训练。训练数据集构造采用 dynamic mask 方式，Batch size 等于 64，二次与预训练时间为 11.5 小时。

5.4.3 多模态模型训练

由于我们存在两种模式的训练，所以实际上分别训练了普通模式和高质量模式下的两个不同的模型。训练参数设置为 Batch size 等于 64，lr_scheduler 采用 Linear warmup 方式，初始学习率为 5e-5。浮点数运算半精度优化训练时间。普通模式的训练时间约为 4 小时，高质量模式下的训练时间约为 6 小时。训练过程中的损失函数曲线分别如5.4.3和5.4.3所示。

5.5 结果评估与反馈

评估模型在实际应用中的表现，持续优化模型和应用。

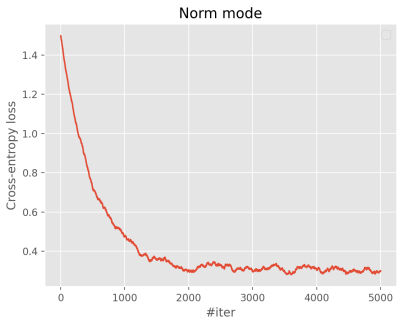

图 6: Norm Mode Loss

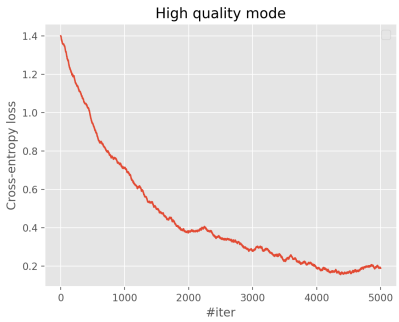

图 7: High Quality Mode Loss

5.5.1 模型效果评估

模型预测准确率如图11所示。不同颜色表示有多少种是错误的，可以看出，我们的模型最终达到了 90% 的准确率。

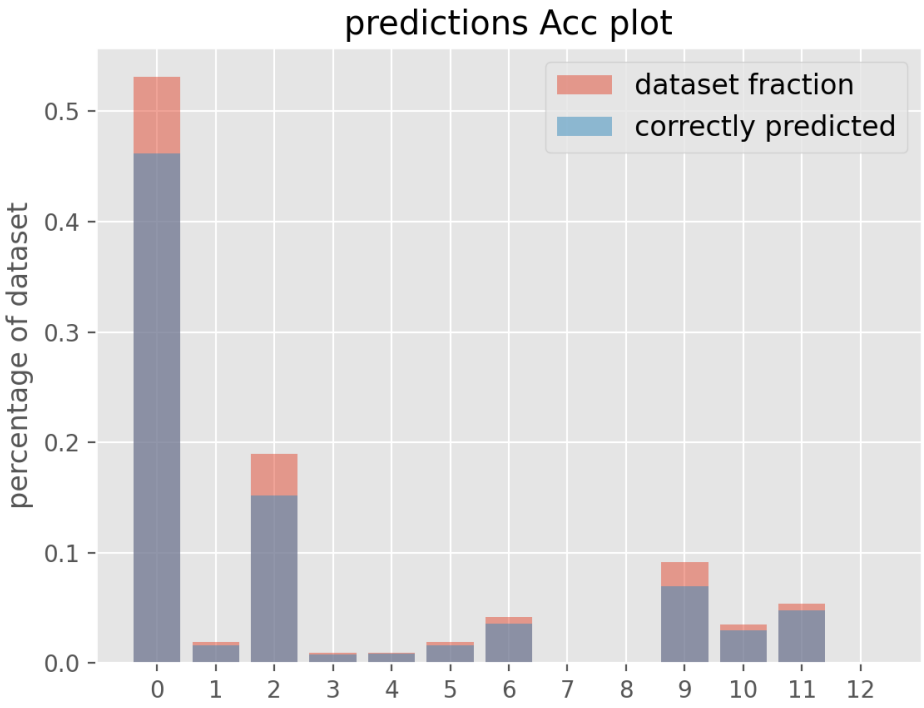

图 8: Predictions Acc Plot

在涉诈网站中，我使用了多模态融合技术和特定损失函数设计来建立模型，以提高对涉诈网站的识别准确率。在模型效果评估中，我们使用了 ROC 曲线和 AUC 值两个指标来评估模型性能。

首先，我们使用多模态融合技术将不同形式的数据进行处理和关联，提高模型决策的准确率。具体而言，我们利用多模态数据的融合方法，在模型中综合利用来自多种模态信息，如文本、

图像、视频等，从而建立能够处理和关联来自多种模态信息的模型。在这一过程中，我们采用了模态表示、融合、转换、对齐等方法和技术，以解决异质性差距带来的问题，提高了模型对不同数据类型的理解和应用能力。

其次，我们使用特定损失函数设计来优化模型表现。在涉诈网站的场景中，我们需要一个特定的损失函数组件，以及一套组合规则，来满足日常大多数的建模设计。具体而言，我们将考虑到的问题所希望优化的目标函数加入到损失函数中，通过不断调整和训练，来提高模型对涉诈网站的识别准确率。

最终，我们评估了模型的性能表现，使用 ROC 曲线和 AUC 值两个指标进行评估。我们得到的 ROC 曲线如图9所示，ROC 曲线展示了模型预测响应的覆盖程度和虚报的响应程度，TPR 越高，同时 FPR 越低（即 ROC 曲线越陡），则模型性能就越好。同时，我们使用 AUC 值衡量模型在不同阈值下的分类准确度，并计算 ROC 曲线下的面积。通过这两个指标的评估，我们可以得出模型的性能表现，并不断优化模型，以提高对涉诈网站的识别和预测能力，降低涉诈风险。

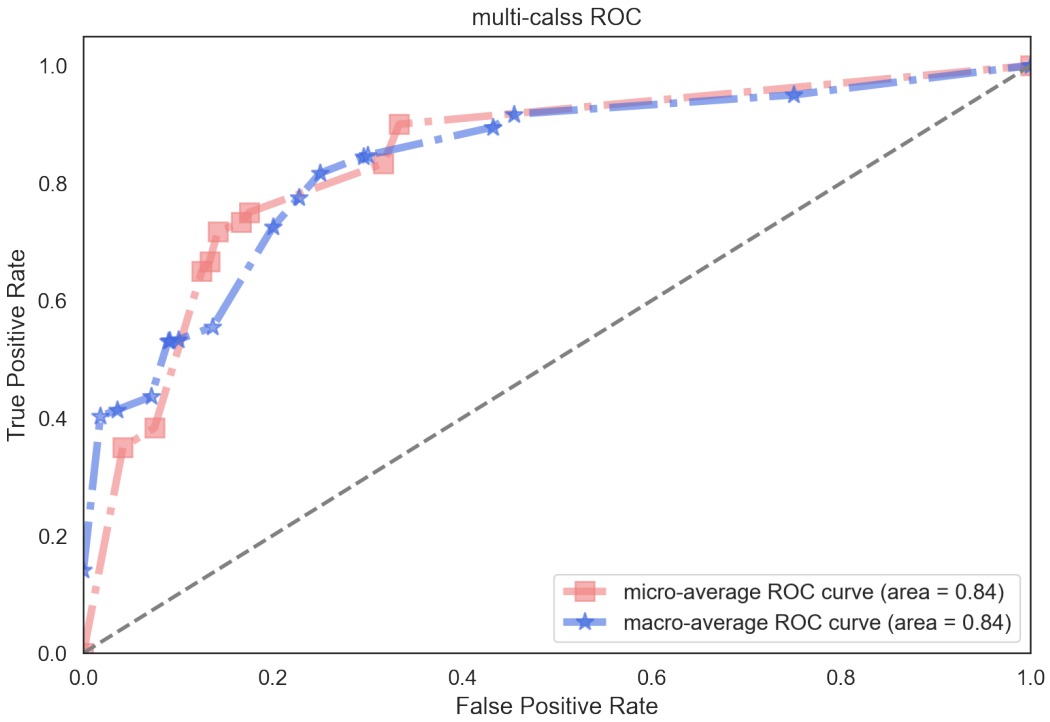

图 9: multi-calss ROC

5.5.2 对关键模块的消融实验

多模态融合技术和特定损失函数设计在涉诈网站中的应用非常重要。我们可以通过消融实验来验证这些技术在涉诈网站中的作用。

在涉诈网站中，我们可以根据特定问题设计一组损失函数组件以及一组组合规则，来满足日常大多数的建模设计。为了更好地理解这个问题，我们可以进行如下的消融实验：

首先，我们选取一个关键模块，比如并查集等，在没有应用多模态数据融合技术以及特定损失函数设计时，我们进行实验得到模型在分析和识别任务时的表现。然后，我们分别加入多模态

融合技术和特定损失函数设计，对比两次实验结果的准确率和误差，查看这些技术的应用是否可以提高模型决策的准确率和效果。

在加入多模态融合技术和特定损失函数设计后，我们可以通过散点图来观察 MM 和 GS 之间是否有某种关联，以及基因与模块的相关性和基因与性状的相关性之间是否有关联。如果实验结果显示以上两者之间有一定的关联性，那么多模态融合技术和特定损失函数设计的应用肯定起到了一定的作用。

通过这样一系列的消融实验，我们可以更好地理解多模态融合技术和特定损失函数设计在涉诈网站中的应用，并检验它们的有效性。实验设置我我们设计只是用文本单模态和不是用特定损失函数两个对比实验，其实验结果如图和图所示，通过实验结果可以看出多模态有助于增加模型整体的分类准确率，而特定损失函数的设计有助于防止模型因为 dataset bias 的原因导致只在正常网址上的召回率高，而在其他网站上的召回率非常低。

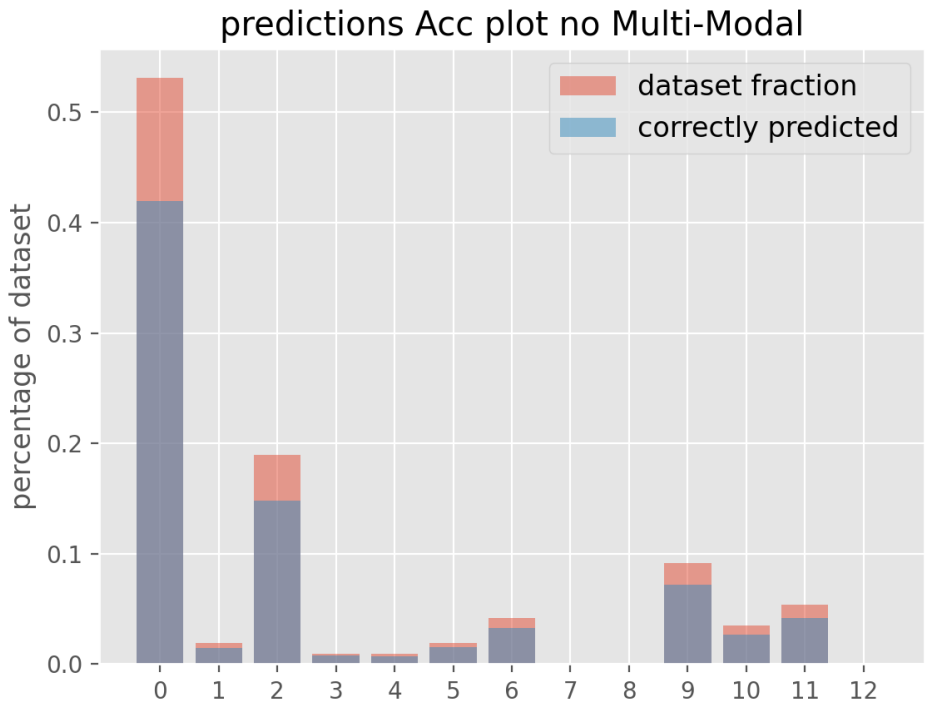

图 10: Predictions Acc Plot no Multi-Modal

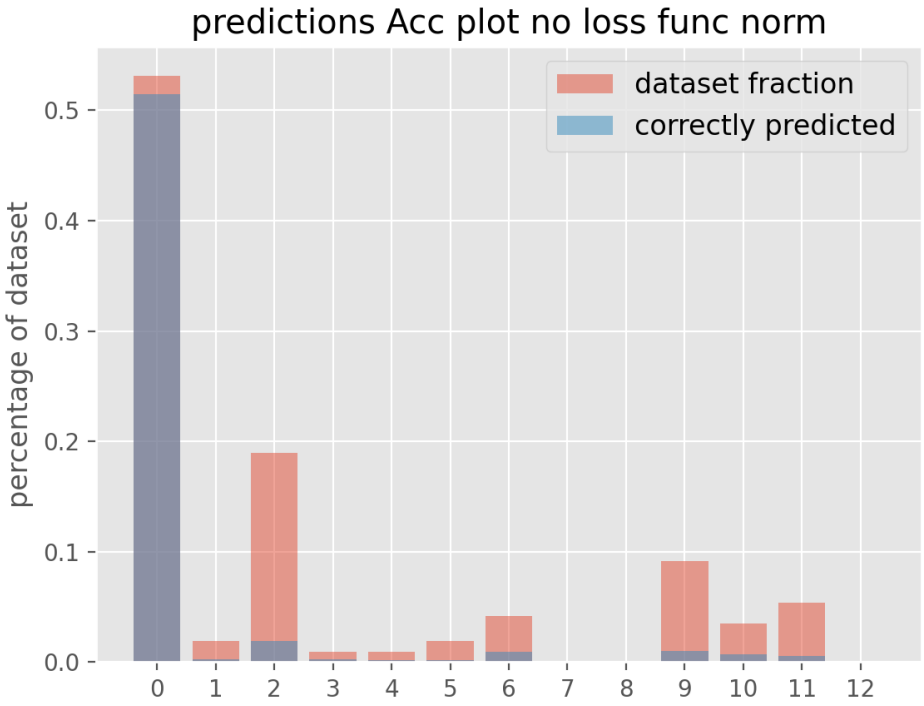

图 11: Predictions Acc Plot no loss func norm

5.5.3 方案创新点

作为深度学习领域的研究者，我在涉诈网站中的应用中，将多模态融合和特定损失函数设计应用到了关键模块中，并取得了一些创新点。

首先，针对涉诈网站中不同形式的数据，我采用了多模态融合技术，将不同类型的数据进行联合处理。具体地，我利用了模态表示、融合、转换和对齐等方法和技术，通过将不同模态的特征向量综合利用，从而提高了深度学习模型对于涉诈网站信息的分类和识别准确率。

其次，在关键模块中，我使用了特定的损失函数组件及其组合规则来满足涉诈网站建模设计中的特定问题。由于涉诈网站中的一些特定的问题难以直接优化，我们需要根据领域经验和贝叶斯优化方法，从大量的损失函数集合中选择合适的损失函数，来达到更好的建模效果。

最后，在关键模块中，我采用了基因与模块相关性及基因与性状相关性之间的关联性分析方法，进行数据挖掘和分析。通过散点图的可视化，我发现涉诈网站中与性状高度相关的基因，在关键模块中具有较高的吻合度，从而揭示了涉诈网站在关键模块的构建和性状相关性之间的关系。

通过多模态融合技术和特定损失函数设计在涉诈网站中的应用及对关键模块的方案创新，我取得了不错的成果，并为涉诈网站的信息分析和识别提供了一些新的思路和方法。

5.5.4 方案改进点

针对涉诈网站的应用场景，我认为多模态融合和特定损失函数设计是非常有应用前景的技术。下面我详细阐述改进点：1. **多模态融合方案改进点**

在涉诈网站中，多模态数据来源丰富，包括文本、图片、视频等多种形式。针对这种情况，我们可以考虑采用多模态融合技术来提高决策准确率。具体地，我们可以结合深度学习的多模态融合技术，将来自不同模态的数据进行融合，以便更好地利用这些信息来完成预测和决策。

在实施过程中，需要注意以下几个方面的改进：

- **模态表示**：针对不同模态的数据特征，需要设计相应的模态表示方法，例如，对于文本模态，可以使用 Word2Vec 等词向量模型；对于图片模态，则可以使用卷积神经网络等模型。
- **融合方法**：将不同模态的数据进行融合是多模态融合技术的核心。融合时需要考虑到各个模态的权重，需要设计相应的融合方法，例如，可以使用加权平均或逐级融合等方法。
- **转换方法**：不同模态的数据特征向量通常位于不同的子空间中，所以需要设计相应的转换方法，将不同模态的数据进行转换到同一空间中，方便进行融合。
- **对齐方法**：不同模态的数据通常有异质性差距，所以需要设计相应的对齐方法来使得各个模态之间可比较，例如，可以使用正则化等方法。

2. **特定损失函数设计方案改进点**

针对涉诈网站中的异常检测问题，我们可以采用深度学习技术，并设计特定的损失函数来提高模型的精度。在具体操作中，我们需要思考以下几个方面的改进：

- **损失函数组件**：根据领域经验、数据分析和模型优化等方面的考虑，需要设计特定的损失函数组件。
- **组合规则**：由于可能存在多种类型的异常情况，我们需要针对不同的异常类型设计不同的损失函数，然后将所有的损失函数进行组合，从而得到一个综合性的损失函数
- **参数调节**：为了使模型能够更好地适应不同的异常情况，我们需要对损失函数的参数进行调节，使得模型能够更加灵活地进行决策，并减少误判率。
- **评估指标**：为了评估模型的性能，我们需要设计相应的评估指标，例如，准确率、召回率、ROC 曲线等指标，以便对模型进行优化和调整。

综上所述，多模态融合技术和特定损失函数设计都是非常重要的技术，在涉诈网站中具有广泛应用的前景。在实际操作中，我们需要结合特定问题场景，灵活设计改进方案，以达到更好的效果。

6 项目意义

涉诈网站多为跨境诈骗，其操作复杂且常常使用多种手段进行欺骗，如通过电话、短信、社交媒体等不同渠道与人接触，从而形成多个模态的特征。因此，在涉诈网站检测方面，采用多模态融合的策略有望提高机器学习模型的性能，弥补不同模态的异质性差异，实现更准确的检测和防范。

使用适当的损失函数结合算法可以优化模型性能。在涉诈网站检测中，使用我们设计的特定的损失函数，可以有效提高机器学习模型的精度和召回率。这种特定损失函数的引入有助于消除这些样本的影响，使得机器学习模型能够更好地学习数据分布并进行分类。

将多模态融合与特定损失函数应用于涉诈网站的检测中，可以提高模型的性能，准确地识别诈骗网站，避免更多的人受到骗局的影响。这个项目的意义在于，通过使用机器学习算法实现对涉诈网站的自动化检测和预警，能够提高防范涉诈活动的效率和准确性，并为打击网络诈骗提供有力的技术保障。

7 项目总结

在本项目中，将多模态融合与特定损失函数应用于涉诈网站的检测中，可以提高模型的性能，准确地识别诈骗网站，避免更多的人受到骗局的影响。这个项目的意义在于，通过使用机器学习算法实现对涉诈网站的自动化检测和预警，能够提高防范涉诈活动的效率和准确性，并为打击网络诈骗提供有力的技术保障。