安卓应用隐私合规性检测系统的设计与实现

摘　要

在App违法违规收集和使用个人信息的现象频繁发生、隐私保护领域法律法规接连出台、应用隐私合规检测困难重重的背景下，为保护公民隐私、促进法律法规落实，设计实现隐私合规性检测系统意义重大。本研究以隐私政策为抓手，通过文献回顾、数据标注、模型训练、系统研发等方式，成功研发了安卓应用隐私合规检测系统。

通过深入研究隐私政策领域法律法规、系统梳理关键要素，本研究构建了涵盖隐私政策核心要求的评价指标体系，共包含7个一级指标和21个二级指标。基于指标体系，提出了隐私政策的量化评估方法。研究还制定了中文隐私政策文本标注手册，对130篇隐私政策中的一万七千余条文本进行多层级标注，构建了高质量的中文隐私政策数据集，为后续隐私政策分析、模型训练和算法优化提供了可靠基础。本研究基于BERT和TextCNN两种网络结构，成功开发出一种高效的中文隐私政策文本分类模型，称为BTC模型。该模型能够准确地对隐私政策进行分类和归类，为隐私政策的自动处理和分析提供了有力支持。基于指标体系和分类模型，本研究开发了一款安卓应用隐私合规检测系统，该系统能够自动化地检测安卓应用的隐私政策合规性。

总的来说，本研究是对中文隐私政策合规性检测领域的一次有价值的探索，有效利用科学技术解决社会治理问题，为增强我国个人隐私保护领域执法、监督能力做出了贡献，对提升用户的个人信息安全水平和对移动应用的信任感具有重要意义。

关键词：中文隐私政策；合规性；数据集构建；BERT；TextCNN

Design and Implementation of Android Application Privacy Compliance Detection System

Abstract

Given the prevalence of illegal collection and use of personal information by mobile apps, increasing privacy laws, and the challenges in app privacy compliance checks, developing a privacy compliance detection system is crucial. This study successfully created an Android app privacy compliance detection system by focusing on privacy policies and conducting literature review, data annotation, model training, and system development.

By conducting comprehensive research on privacy policy laws and identifying key elements, the study established an evaluation index system with 7 primary indicators and 21 secondary indicators, covering the core requirements of privacy policies. A quantitative evaluation method was proposed based on this index system. Chinese privacy policy texts were annotated, and a Chinese privacy policy text classification model named BTC was developed by combining BERT and TextCNN network structures. This model accurately classifies and categorizes privacy policies, providing support for automated processing and analysis. An Android app privacy compliance detection system was built upon the index system and classification model, enabling automatic detection of privacy policy compliance in Android apps.

In conclusion, this study is a valuable exploration in the field of Chinese privacy policy compliance detection. It addresses social governance issues, enhances law enforcement and supervision in personal privacy protection, and improves users' personal information security and trust in mobile applications through scientific and technological approaches.

Key Words: Chinese privacy policy; Compliance; Dataset construction; BERT; TextCNN

目 录

第1章 绪论

1.1 研究背景及意义

新一代信息技术蓬勃发展，改变了人们生活的方方面面，开发和利用信息科学与技术更是成为了提升国家和社会治理效率的重要抓手，成为了实现国家治理体系和治理能力现代化的关键契机。习近平总书记强调，要着力推进社会治理系统化、科学化、智能化、法治化。因此，找准当前社会治理的需求点，通过开发智能辅助工具提升治理质效已经是一个新兴的重要议题。

近年来，App违法违规收集、使用个人信息，侵犯公民隐私权的现象屡见不鲜，引发了人们的广泛关注，我国执法机关也开始着力整治这一问题。据相关数据统计，87% 的移动APP存在隐私违规问题；2020年“315晚会”曝光50款APP存在窃取隐私问题；成都多个教育APP被投诉出现疑似隐私泄露的问题；教育旅游类APP因过度采集网民个人隐私信息，成为APP隐私违规的重灾区。2022年11月3日，国家网信办依法集中查处一批侵犯个人信息合法权益的违法违规App，对其中的55款App进行下架处理，其余80款App勒令限期整改。在此前，滴滴出行等著名App也曾因为过度收集用户个人隐私被下架。据工信部统计，2018年时我国有App共421万款，到2022年8月，数量已经下降为了232万款，可见我国近年对于应用隐私合规的监管愈发严格。

虽然我国当前已开始对App收集和处理个人信息进行规制，但是在实践中仍然面临着一些问题。即便我国App数量在近几年内已经出现了明显的下滑，但其总数仍是百万级别，而App收集了哪些个人信息、如何使用个人信息大部分情况下处于不透明的软件黑箱中。如果使用人工对于App收集处理个人信息的行为进行审核，将消耗巨额的人力和时间成本。

隐私政策是对于App收集和处理个人信息进行规制和检测的重要切入点。隐私政策是一份有关企业等信息收集方就如何收集个人信息所发布的声明，信息收集方会在隐私权政策中提到自己收集、存储、使用、披露和管理客户数据的政策[1][2][3]。信息收集方在声明中会告知客户哪些具体信息会被收集，以及这些信息是否会被保密，是否会与合作伙伴共享，或出售给其他公司或企业。根据《App违法违规收集使用个人信息行为认定方法》中的规定，我国的App都需要具有并按照国家法律规定的相关内容相应的制定隐私政策[4]。

隐私政策合规性主要分为两个方面。一方面是隐私政策的内容特性，即是指组织或公司在处理个人数据时，需要遵守相关法律，以确保其隐私政策符合适用法律和法规的要求，并为个人数据主体提供透明和清晰的信息，使其能够了解其个人数据如何被收集、使用、共享和保护[5]。另一方面则是隐私政策的形式特性，我国已明确规定App的隐私政策应单独成文、易于阅读、易于访问，并对敏感信息应给出显著标识[6]。因此，隐私政策合规性是指其内容和形式上符合法律法规要求的程度。

即使当前大部分App都都已经制定了隐私政策，但是隐私政策不合规的现象仍然频出。例如，工信部自2019年以来已经公布了29批侵害用户权益行为的APP名单，每一批均有软件因“应用分发平台上的App信息明示不到位”、“收集个人信息明示、告知不到位”、“违规收集个人信息”和“超范围收集个人信息”等直接与隐私政策合规性相关的问题被通报[7][8]。

如何有效检测App的隐私政策不合规问题成为了重大挑战。对于监管机关而言，我国当前App数量达数百万，如何对于如此巨量的应用进行隐私合规审查成为了一个难题；对于相关企业而言，法律法规规定纷繁复杂，即便专业法律人士也很难完全厘清所有细节，因此企业对其研发的软件进行使用前的合规性审查困难重重；对于产品用户来讲，App隐私政策文件冗长、晦涩，少有用户会详细阅读并分析考量[9]，根据统计，大约77.8%的用户在安装App时很少或从未阅读过隐私协议[10]。

针对于上述问题，设计与开发自动化的隐私合规检测系统将成为一个有力的解决方案[11]。自动化的隐私合规检测系统往往依托于自然语言处理和深度学习技术，通过理解隐私政策中的语义信息，分析隐私政策是否符合法律法规规定[12][13]。当前已经有一些企业和政府机关开展了相关的产品设计与开发实践，但尚处于初期阶段，检测系统、平台的检测精确度、应用场景范畴和使用对象范畴都存在着提升和扩展的空间。因此深入分析当前隐私政策自动化检测中存在的不足，并持续推进隐私政策合规检测系统研发十分重要。

本研究特别关注了安卓应用隐私政策合规性的检测问题。最近的数据显示，安卓占据71.77％的全球移动操作系统市场份额，而iOS占据了移动操作系统市场的27.6％，全球有超过25亿活跃的安卓用户，分布在190多个国家[14]。安卓应用拥有着最大用户群体和市场占有率，同时面临着较多的个人信息保护问题，应当首先得到关注[15]。

总之，本研究在国家大力推动科技社会治理的背景下，进行安卓应用隐私合规

性检测系统的设计与实现，以应用隐私政策为抓手，实现隐私合规的自动化检测，为监管机关、技术企业和App用户提供隐私问题的解决方案。

1.2 研究目标和内容

本研究旨在研发一种自动化检测系统，以检测App是否违反隐私政策，从而帮助司法机关、企业及产品用户有效地发现和防范App隐私政策合规性不足的为题。具体来说，研究内容包括：一是系统梳理当前隐私政策合规性领域的相关文献和法律法规规定，建立隐私政策合规性评价指标；二是收集和整理大量App的隐私政策数据，并对其进行分类、标注和分析，建立一个高质量中文安卓应用隐私政策数据集；三是设计一种能够有效识别和提取隐私政策信息的方法，依托本研究建立的高质量中文安卓应用隐私政策数据集，利用深度学习等技术研发一个隐私政策文本分类模型，并对模型效果进行评估和对比；四是开发一个隐私政策合规性自动化检测分析系统，对App隐私政策进行检测和分析，并对系统进行测试和评估，比较其与人工审核的准确率、效率和成本，以验证该系统的实用性和有效性；五是探讨本文提出的安卓应用隐私合规性自动化检测系统在司法机关、企业以及产品用户中的应用前景和推广方案，为进一步提高App隐私保护水平、维护公民隐私权益提供有益的参考和支持。

1.3 研究创新点

本文针对上述研究空白开展了深入研究，主要形成了以下三方面创新点：

一是建立了隐私政策合规性评价指标。当前世界范围内对于隐私政策合规性评价指标构建都有一定欠缺，这是因为全球各国从2018年前后才开始广泛关注个人信息保护领域的相关问题，我国的隐私政策相关法律法规规定也主要都是于近五年内出台。同时由于不断有新的法律法规规定出台，评价体系也要据此不断更新，并适应自动化检测的要求。通过深入阅读相关文献，系统梳理隐私政策领域的法律法规规定，本研究建立了包含7个一级指标和21个二级指标的隐私政策评价方法，并且给出了包括虚假性评估、完整性得分和内容分布得分在内的3类量化评估手段，有效助力隐私政策的合规性评价。

二是构建了高质量标注的中文隐私政策数据集。当前隐私政策领域已经有数个被广泛应用的开源英文隐私政策数据集，对于英文隐私政策的分析和检测提供了巨大帮助。然而，虽然中文隐私政策数量众多，但少有高质量标注过的中文隐私政策

数据集，少量这一领域研究者开发的中文隐私政策数据集存在着局限性强、没有进行开源等根本性问题，难以支撑中文隐私政策分析和检测研究的开展。本研究依托隐私政策合规性评价指标，形成了中文隐私政策文本标注手册。在进行严格的数据清洗的基础上，对于130篇隐私政策中的一万七千余条隐私政策文本进行多层级标注，形成了高质量的中文隐私政策文本数据集。

三是研发了更高效的隐私政策文本分类模型。隐私政策文本分类是开展隐私政策合规性检测及自动化检测系统研发的基础性工作，但当前领域内对于分类模型研究较少，且效果仍有提升空间。本研究采用近年来在文本分类任务上效果卓著的Text-CNN、BERT等深度学习模型，进行了大量实验和分析，最终成功高效的中文隐私政策文本分类模型。

1.4 研究框架

本研究通过系统梳理文献，深入研读法律法规规定，建立隐私政策合规性评价指标；经过数据收集、清洗和标注，构建中文隐私政策高质量数据集；通过实验验证，研发了高效的中文隐私政策文本分类模型；最终整合研究成果，开发了中文安卓隐私政策合规性自动化检测系统。本文对研究内容进行了规范的安排，主要框架可分为以下几个部分：

第一章是绪论。从我国大力推动科技社会治理的背景讲起，指明研究意义、研究目的研究目的和内容、研究创新点以及研究框架。

第二章是文献综述。根据第一章指出的研究问题，开展文献调研，梳理归纳国内外相关研究和法律法规规定情况，奠定研究的理论基础。对于隐私政策、合规性、完整性等关键概念进行界定，介绍当前隐私政策合规性检测领域的主要工作情况和研究基础。

第三章是安卓应用中文隐私政策数据集构建与分析。这一章节主要介绍本研究构建的中文安卓隐私政策数据集，系统描述了数据集标签体系建立、数据标注过程和数据集描述性统计情况。

第四章是中文隐私政策文本分类模型研究。通过系统分析Word-vec、Text-CNN和BERT等在文本分类领域效果卓著的深度学习相关技术，建立了中文隐私政策文本分类模型，并进行实验验证，对比分析了分类模型的效果。

第五章是安卓应用隐私政策合规性检测系统的设计与实现。本章节整合前章研究成果，介绍了包括虚假性评估、完整性得分和内容分布得分在内的三类量化评估方法，描述了安卓应用隐私政策合规性检测系统的设计方案，阐述了系统的实现过程，并对系统进行测试，验证了系统的应用效果。

第六章是总结与展望。这一章节对于研究进行整体回顾和系统梳理，指出了研究中存在的不足和局限性，点明研究的主要贡献，并对于研究成果的应用和进一步发展做出了展望。

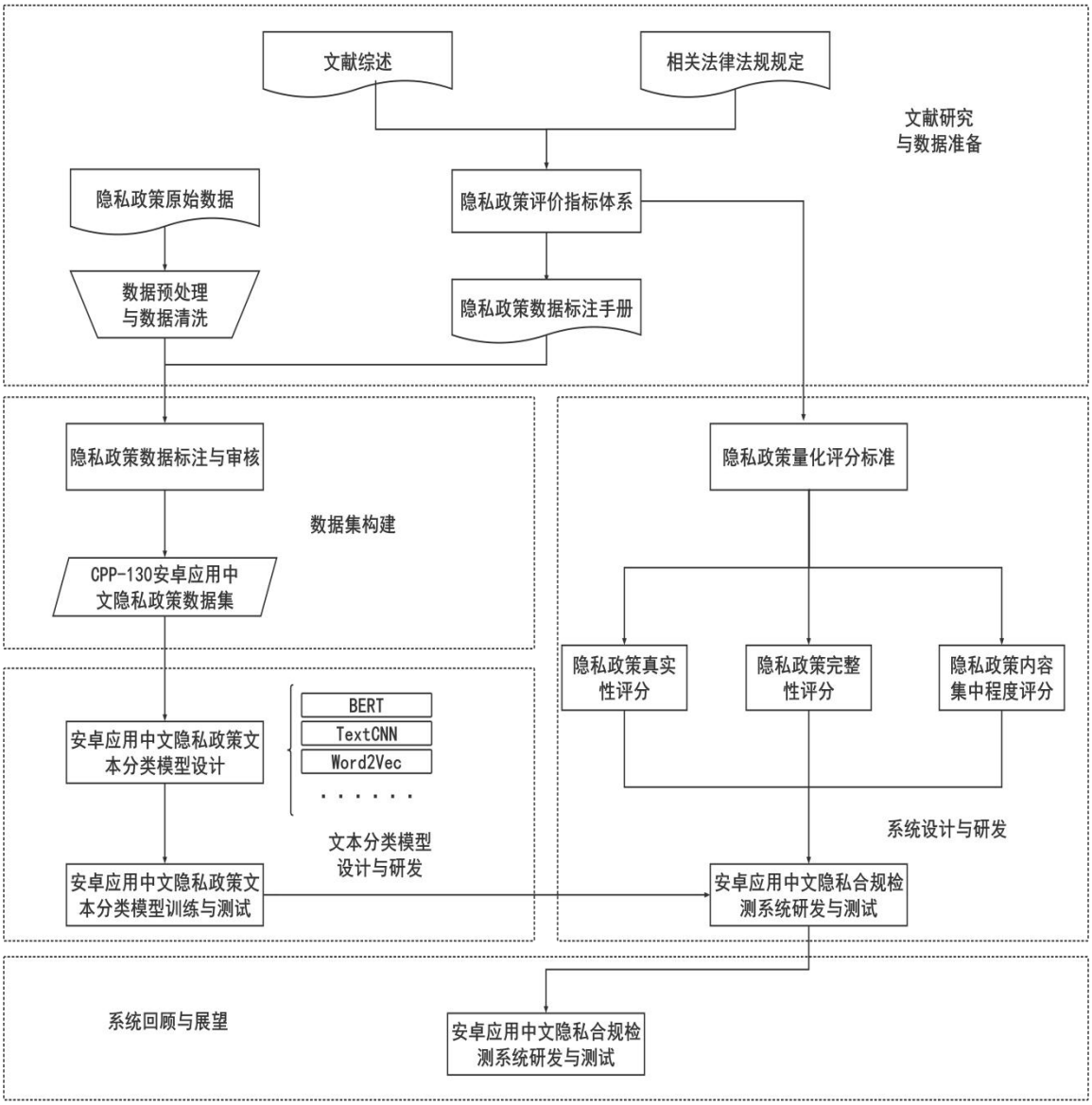

图1-1 研究框架与技术路线图

第2章 研究现状

本章将着重分析和概述有关安卓应用隐私政策合规性的相关研究和发展。随着移动应用的普及和用户对于隐私保护的关注度增加，如何有效检测应用程序的隐私政策合规性问题成为了一个重要的研究领域。本章将从隐私政策合规性评估、中英文隐私政策数据、国内外隐私政策相关法律法规以及隐私政策文本分类模型研究这四个方面综述当前隐私合规检测领域的的主要工作情况和研究基础。

2.1 隐私政策合规性评估

隐私政策合规性评估方法种类众多，具体包括真实性、完整性和可读性等[16][17]。

2.1.1 隐私政策真实性

隐私政策真实性并不是一个得到广泛研究的概念，与之相近的概念有隐私政策的有效性，同时一些学者曾经对于虚假隐私政策开展检测，同样探讨的是隐私政策真实性的问题。朱璋颖等（2020）曾对虚假隐私政策开展研究，并给出了虚假隐私政策的评判公式，主要根据隐私政策中数据实践相关表述所占比例评判隐私政策是否真实[18]。姜盼盼等（2019）从法律角度指出，要确保隐私政策链接内容真实有效，跳转方式直接，显示位置公开[18]。隐私政策主要用于描述应用对于用户个人信息的处理方式，其真实性是隐私政策存在的基础。本文在假设应用中存在的隐私政策链接都是真实有效的基础上，将隐私政策真实性定义为隐私政策是否主要描述个人信息相关的数据实践。

2.1.2 隐私政策完整性

隐私政策的完整性是指隐私政策对标相关法律法规政策的完备和全面的程度。例如《中华人民共和国个人信息保护法》[19]规定：“个人信息的处理包括个人信息的收集、存储、使用、加工、传输、提供、公开、删除等。”在隐私政策中，有些App没有完整包含法律规定的内容即属于完整性不足。

国内研究方面，张艳丰等（2021）在中国法律体系背景下，构建了一个用户感知的个人信息保护政策合规性测度体系，包括个人信息保护政策与标准规范的一致性、文本完整性、位置显著性和内容可读性等四个维度，使用灰色加权关联分析方法，对移动阅读类应用进行了关联度计算和排序[20]。朱璋颖等（2020）以《信息安

全技术个人信息安全规范》（下简称《规范》）为主要参考建立指标体系，通过支持向量机方法进行分类，并建立公式评估隐私政策完整性[18]。赵波等（2020）同样依托《规范》分析国内当前监管条件下移动互联网应用程序个人信息安全的相关要求，结合Android应用程序个人信息安全检测技术，利用层次分析法，提出了Android应用程序个人信息安全量化评估模型[21]。

国外研究方面，Torre等人（2021，2020）基于GDPR构建相关条款，并使用自然语言处理和监督式机器学习开发自动化隐私政策完整性检验软件[22][23]。Mousavi等人（2020）提出了三种不同的模型，用于使用监督机器学习将隐私政策的段落分类为预定义的类别[24]。Bhatia等人（2016）开发一个半自动化框架，通过众包和NLP从隐私政策中提取隐私目标[25]。Fan等人（2020）检测移动健康应用程序是否符合 GDPR 完整性，定义了六类隐私相关信息，使用基于机器学习的二元分类器应用于给定策略中句子的词袋表示形式，以预测隐私政策是否完整[26]。

总的来说，当前隐私完整性检测的主要是技术是先建立一个用于测评政策文件完整性的指标体系，即根据相应的法律法规，列出一份隐私政策里面需要包含的内容，更进一步的还可以给这些内容赋予权重；随后，利用自然语言处理技术和机器学习模型，对于未分类的隐私政策语句进行分类，类别即为上一步中建立的指标体系，将语句与对应类别的评价指标进行比对，根据比对结果即可完成完整性检测。

2.1.3 隐私政策可读性

隐私政策的可读性是指隐私政策对于用户来说容易阅读和容易理解的程度。隐私政策参考相关法律法规文件制定，往往具有着冗长、晦涩的特点。早有研究在2008年就指出，如果一个用户要阅读在互联网上访问的每一项服务的隐私政策，平均每年需要244 h，因为鲜少有用户真的会去阅读隐私政策。另外，大量的法律术语，对于缺乏专业知识的普通用户外也存在着专业壁垒。

陈世敏（1971）最早提出了“陈世敏公式”用于判断中文可读性[27]，在隐私政策可读性分析领域同样被广泛应用（苗慧等，2021；秦克飞等，2019）[28][29]。朱侯等人（2018）提出字体、行距、小标题数等10个指标用于评估隐私政策可读性[30]。

综合文献，目前可读性检测主要采用的是公式法，根据不同的语言种类，要开发不同的可读性检测公式，国内已经有不少学者开展了中文可读性公式的开发，其中主要包含的可读性影响因素涉及平均笔画数、难字或完全对称字率、特殊词或难

词率、句均字或词数等。然而，当前专门针对于中文隐私政策可读性的研究并不丰富，仍然存在着较多的研究空白。

2.2 中英文隐私政策数据集

2.2.1 英文隐私政策数据集

为了研究隐私政策的内容和数据实践，已经有学者提出了一批较为成熟的英文隐私政策数据集。最为经典的是Wilson等人于2016年提出的OPP-115数据集（Online Privacy Policies, set of 115）[31]，该数据集包含了115篇隐私政策，由熟练的注释者（法学院学生）手动进行了注释。在注释过程中，注释者按照10个高级别类别将段落大小的隐私政策段落进行了标注，然后选取了每个段落中的部分内容，并使用属性-值对进行了注释。这些属性-值对包括了20个不同的属性和138个不同的值。在所有的值中，有122个值具有超过20个标签。此外，数据集中还展示了每个段落中必须包含的属性。图2-1中展示了部分类别和属性的情况，该数据集的详细注释和内容为研究者提供了探究隐私政策和数据实践的重要资源。

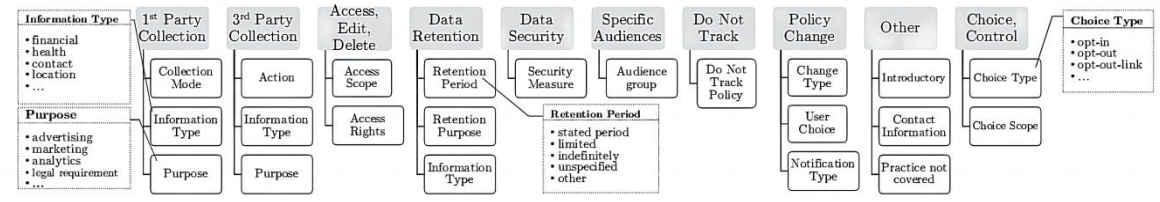

图2-1 Wilson等人的的隐私分类[31]

OPP-115主要关注的是在线网站的隐私政策，英文隐私政策领域的研究者后续又基于OPP-115进一步研发了新的一系列数据集[32][33]，关注了隐私政策中其他重要的方面。Opt-out Choice Dataset介绍了一种使用网站隐私政策文本来训练机器学习和自然语言处理模型，以识别选择（例如，退出行为广告）的方法，并创建了一个网站隐私政策文本语料库[34]。MAPS Policies数据集包含441,626个隐私政策的URL，隐私政策是由Mobile App Privacy System（MAPS）在2018年4月至5月期间对Google Play商店应用程序进行分析时发现的。APP-350 Co数据集，由350个Android应用隐私政策组成，这些隐私政策已经被标注了隐私实践，即可能涉及隐私的行为[35]。Poplavska等人（2020）根据GDPR在OPP-115的基础上进行了调整，建立了数据集与最新法律法规之间的联系[36]。隐私法律语料库（2022）则提出了一个包含1043个隐私法律、

法规和指南的数据集，涵盖了世界182个司法管辖区，数据集中的文档以pdf和txt两种文件格式提供，还提供了部分多语种版本[37]。

总的来说，当前英文领域的隐私政策数据集已经较为丰富，既有大量隐私政策的原始数据，也有从各种角度进行高质量标注的数据。同时，已经有大量研究者利用这些数据集开展研究并且得到了广泛认可和应用。

2.2.2 中文隐私政策数据集

相较于英文领域隐私政策数据集的蓬勃发展，中文隐私政策数据集的研发仍然处于较为初级的阶段。

事实上，目前几乎没有优质的开源中文隐私政策数据集，仅有部分研究者按照其个性化的研究目标进行了数据标注供自己的研究使用。朱璋颖等（2020）在开展中文隐私政策文本分类模型研究时，通过众包方式标注了Chinese-OPP-100，主要参照OPP-115的标注方法[18]。赵杨（2022）等在探究基于机器学习的医疗健康APP 隐私政策合规性研究时，建立了评价指标体系并进行数据标注[38]。朱侯等人（2023）在进行基于BERT文本分类模型的APP隐私政策完整性评价研究时，[39]。这些主要都参考了成熟的英文数据集的构建方式，质量水平尚可，然而一方面没有进行开源，无法经过更多研究者的检验，也无法得到广泛应用，另一方面由于此类数据集标注时往往具有较强的针对性，只能应用于研究本身的内容，无法进行泛化和扩展。

当前中文隐私政策唯一的开源高质量数据集是Zhao等人于2022年最新发表的CA4P-483数据集[40]。该数据集包含 483 条中文 Android 应用程序隐私政策、超过11K个句子和 52K 个细粒度注释。然而，由于CA4P-483数据集刚刚发表，尚未有学者利用该数据集进行研究并发表。同时，CA4P-483数据集主要对于7类数据实践行为进行细粒度标注，适用于命名实体识别和情感分析等任务，对于文本分类任务则无法适用。此外，该数据集的指标体系建立过程相对简单，其严谨性和全面性仍然有待考量。

总的来说，当前中文隐私政策领域缺乏优质的数据集，少量现存数据集缺乏开源共享、泛化能力较差、质量参差不一且应用领域受限，因此参照优秀的英文隐私政策数据集的研发经验开发高质量的中文隐私政策数据集十分重要。

2.3 国内外隐私政策相关法律法规规定

2.3.1 国外隐私政策相关法律法规规定

欧洲当前App隐私政策相关法律主要是2018年公布的《欧盟通用数据保护条例（GDPR）》[1]。该法规于2018年5月25日生效。它旨在加强个人数据保护，赋予个人更多的控制权和权利，同时强调企业的责任和义务。GDPR涵盖了在欧盟范围内处理个人数据的所有公司和组织，无论其是否位于欧盟境内，只要其处理欧盟居民的个人数据即受到GDPR的监管。GDPR要求企业和组织要采取技术和组织措施，确保个人数据的保密性、完整性和可用性，并在出现数据泄露等问题时尽快通知相关方。GDPR还规定了一系列处罚措施，对违反条例的企业和组织进行处罚，包括罚款和声誉损失。

美国的隐私法律比较分散，各州之间的立法和执行方式也有所不同。《加州消费者隐私法》（CCPA）于2020年1月1日生效[41]。该法规定了一系列要求，包括企业要通知消费者他们收集的个人信息，消费者有权要求企业删除其个人信息等。《消费者数据隐私法》（CDPA）[42]，这是弗吉尼亚州于2021年通过的一项隐私保护法案。该法案规定了企业如何收集、使用、存储和披露消费者个人信息，以及消费者如何访问、更正和删除其个人信息的权利。CDPA适用于在弗吉尼亚州境内经营的企业，包括跨州经营的企业，只要它们在弗吉尼亚州收集、处理或持有消费者的个人信息。此外，还有一些行业相关的法律和监管措施，例如医疗保健领域的《健康保险可移植性和责任法案》（HIPAA）[43]，金融服务领域的《格兰-莱奥尔-鲍利法案》（GLBA）[44]，以及儿童在线隐私保护法案（COPPA）等。英国则制定了《数据保护和数字信息法案》。日本、俄罗斯、泰国、新加坡和澳大利亚等国也均有本国的相关法律法规规定。

2.3.2 国内隐私政策相关法律法规规定

为了保护个人隐私和安全，我国已陆续出台相关法规政策，以规范个人信息控制者在信息处理环节中的行为，主要针对信息的收集、存储、使用和共享等行为。这些法规政策的出台旨在统筹发展和安全，推动数据安全建设，加强对公民的个人信息保护水平。在近五年内，如图2-2所示，各层次部门先后颁布了许多法规政策。在2021年之前，我国应用隐私安全主要是以《网络安全法》为核心，辅以其他的规

章和政策体系。2021年后，随着《数据安全法》和《个人信息保护法》的陆续颁布，我国逐渐形成了以《个人信息保护法》为核心的应用隐私安全体系。

图2-2 个人信息保护相关法律法规梳理

如图2-3所示，本文还总结了部分法律法规规定之间的相互引用关系。当前我国的个人信息保护在国家的战略性、前瞻性部署，以及系统性、完整性的定测设计之下，已经逐渐形成了层次鲜明、内容详实的规制体系。既有高位阶的纲领性法律条文，也有低位阶内容详尽的规章制度。在立法上，由全国人大及其常委会制定法律。随后，由国务院出台行政规章，提出法律落实过程中的必要规范性要求。接着，由国务院下属的包括工信部、网信办、公安部、市监局在内的相关机构，进行执法和监督。更进一步的，由中国信息通信研究院、全国信息安全标准化技术委员会、公

安部网络安全保卫局和中国网络安全审查技术与认证中心等行业协会和委员会，制定相关行业标准和规范，帮助企业、机构和个人落实隐私合规要求。

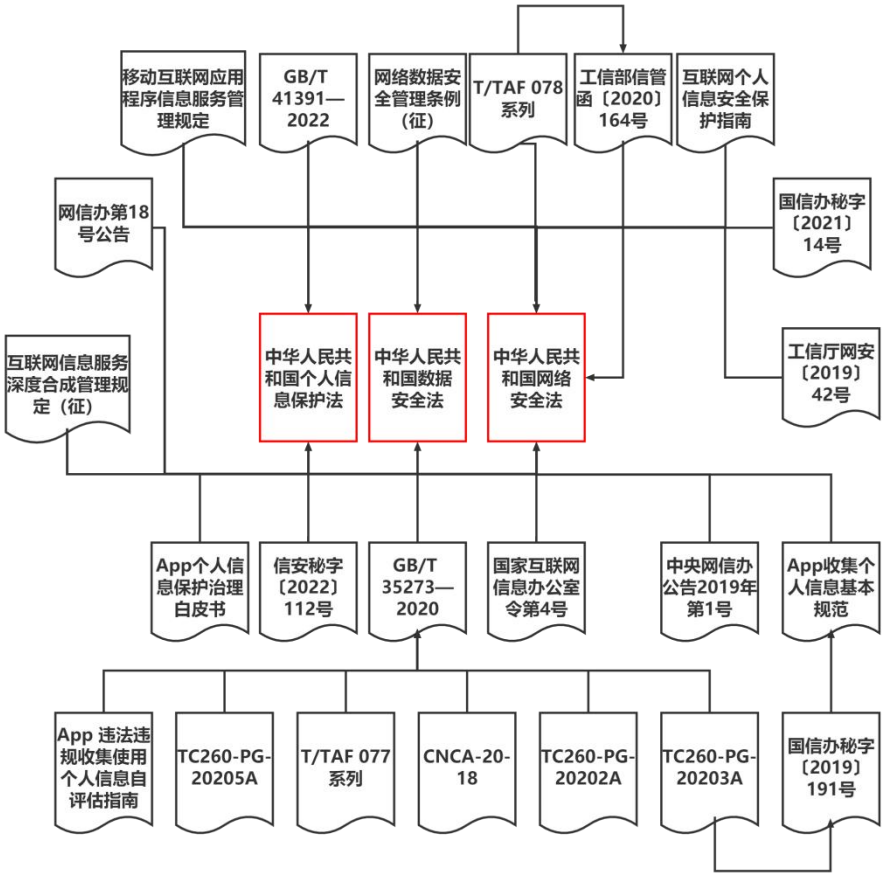

图2-3 法律法规规定之间的相互引用关系

2.4 隐私政策文本分类模型及相关技术

当前已经有一些学者开展了隐私政策文本分类模型及相关技术的研究[45]，有以下几类技术被广泛应用：

支持向量机（SVM）是一种用于分类、回归和异常检测的有监督学习算法[46]。在文本分类任务中，SVM可以利用文本特征构建特征向量，并使用核函数进行高维特征空间的映射。

逻辑回归（LR）是一种简单而常用的分类算法，通常用于解决二分类问题[47]。在文本分类任务中，逻辑回归可以利用文本特征构建特征向量，较好地解决线性可分问题，但对于非线性问题的分类效果较差。

朴素贝叶斯是一种基于贝叶斯定理和特征独立假设的分类算法[48]。在文本分类任务中，朴素贝叶斯可以利用文本特征构建特征向量，并计算每个特征在不同类别下的条件概率。

隐马尔可夫模型（HMM）是一种用于序列数据建模的概率模型[49]。在文本分类任务中，HMM可以将文本看作是由一个潜在的马尔可夫链生成的序列，通过计算给定观测序列的概率，来确定文本所属类别。

长短时记忆网络（LSTM）是一种循环神经网络，用于解决序列数据的建模问题[50]。在文本分类任务中，LSTM可以捕捉文本中的长期依赖关系，并对序列数据进行分类。LSTM的主要优点是能够处理长序列，且能够防止梯度消失或爆炸问题的发生。双向长短时记忆网络（BiLSTM）通过前向和后向的LSTM单元对输入序列进行建模[51]。在文本分类任务中，BiLSTM可以利用文本中的上下文信息，提高模型的准确性。

卷积神经网络（CNN）是一种用于图像识别、自然语言处理等任务的神经网络[52]。在文本分类任务中，CNN可以通过卷积层、池化层和全连接层等结构对文本进行特征提取和分类。CNN可以捕获不同尺度的文本特征，且能够并行处理多个文本样本，因此在文本分类任务中表现优异。Word2vec是一种基于神经网络的词向量表示方法，可以将单词转换成向量，从而实现对单词的语义和语法信息的建模[53]。可以用于提取文本特征，在自然语言处理中得到了广泛的应用。将Word2vec得到的词向量用于文本分类任务，可以得到更好的分类性能。在文本分类中，Word2vec可以将每个单词表示成一个定长的向量，从而将文本表示成一个矩阵，然后使用CNN对其进行分类。使用Word2vec表示文本，可以保留单词之间的语义关系，同时可以解决单词维度过高的问题，从而提高了分类性能。因此，Word2vec与CNN的结合是一种常见的文本分类方法，得到了广泛的应用。

双向编码器表示Transformer（BERT）是一种预训练语言模型，能够在多种自然语言处理任务中取得优异的性能[54]。在文本分类任务中，BERT可以将文本映射为高维向量表示，并利用这些向量进行分类。BERT使用Transformer网络结构进行建模，能够处理长文本序列，且能够利用双向上下文信息进行预测，因此在文本分类任务中表现优异。BERT已成为目前自然语言处理领域的热门算法之一。

总的来说，对于隐私政策文本分类领域，前人已经做了一些探索。然而，在中文隐私政策文本分类模型研究领域仍然存在不足，大部分都是基于英文预料进行训练。此外，使用的方法类别较少，实验验证不够充分，存在着进一步研究的空间。

第3章 安卓应用中文隐私政策数据集构建与分析

本章依托前章中对于我国个人信息保护和隐私政策领域的相关法律法规规定的系统性、深入性梳理结果，建立了中文隐私政策的指标体系。整理了130篇安卓应用中文隐私政策数据，并依托指标体系内容，形成数据标注手册。随后对数据进行清洗和预处理，并参照标注手册，对全部数据进行标注，形成了含有一万七千余条文本数据、三万余个标签的安卓应用中文隐私政策数据集“CPP-130”。最后，对于数据集的基本情况进行了统计和对比分析。

3.1 安卓应用中文隐私政策指标体系

通过系统梳理我国隐私政策保护领域的法律法规规定，本研究主要参考《规范》进行指标体系构建。同时参考了包括《个人信息保护法》、《网络安全法》、《数据安全法》、《移动互联应用程序个人信息保护管理暂行规定》在内的多部法律法规。《规范》是我国当前隐私政策合规领域应用最为广泛、最为权威的规定，详细指明了有关个人信息收集、共享、存储和使用等数据实践行为应当遵守的基本原则和要素。《规范》明确要求了隐私政策应当清楚展示企业、应用或平台方将会如何处理用户个人信息，并详细规定了隐私政策应当如何编写，给出了相关的范例。

依据上述法律和规章，本研究初步建立了一个包括7个一级指标和38个二级指标的指标体系，如图3-1所示。7个一级指标对应的是安卓应用对于用户个人信息和隐私数据进行的数据收集、共享、处理等实践行为的主要类型，其中“其他通用信息”类包含非数据实践相关的隐私政策内容。38个二级指标则建立在7类一级指标之下，对应数据实践行为大类之下的具体内容。法律和规章中对于这些指标内容均做出了明确规定，即隐私政策中应当包含指标体系中的内容。

该指标体系较为系统的反映了当前我国法律法规规定中对于隐私政策的相关要求，若一篇隐私政策涉及指标体系中的指标信息，即可视为该隐私政策的合规性基本达标。

以“与第三方共享、转让、公开”这一一级指标为例。“与第三方共享、转让、公开”是指，第一方（即软件运营者）将用户信息与第三方（包括合作伙伴、其他软件服务商等）分享、转让或公开的行为。其具有11个二级指标，内容和含义如表3-1所示。“与第三方共享、转让、公开”的典型文本有：“我们可能会将您的信息

共享给我们的合作伙伴，包括广告合作伙伴。我们只会共享必要的信息，并且这些合作伙伴将遵守本隐私政策和相关的数据保护法律。

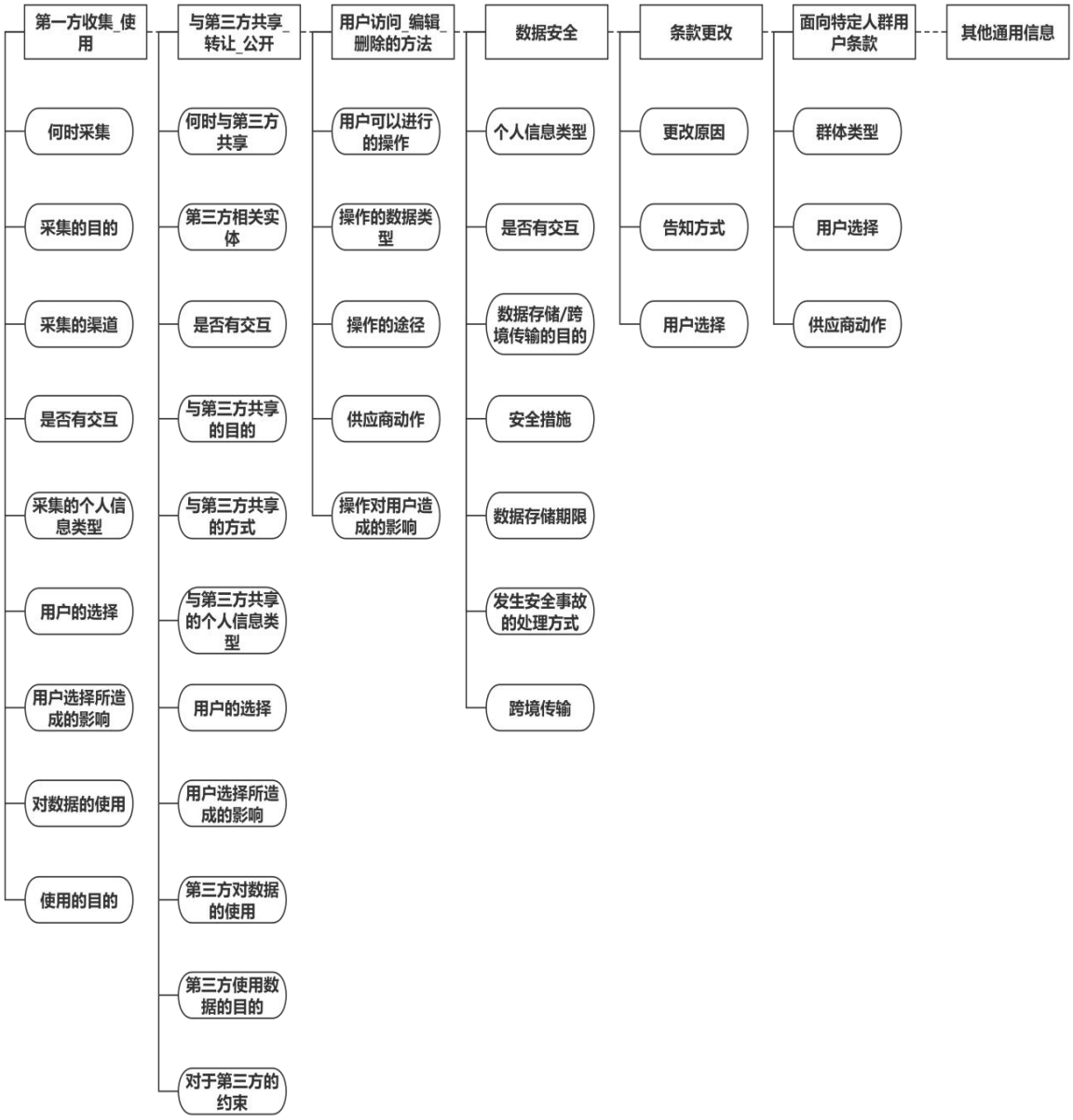

图3-1 安卓应用中文隐私政策多级指标体系

指标体系建立后，安卓应用中文隐私政策数据集标注任务的标签体系相应产生，即指标体系中的7个一级指标对应于标注过程中的7个一级标签类型，38个二级指标对应标注过程中的38个二级标签类型。

表3-1 “与第三方共享、转让、公开”的二级指标内容及释义

二级指标名称	二级指标释义
何时与第三方共享	指第一方何时与第三方共享用户信息，例如在注册、交易、咨询等过程中
第三方相关实体	指与第三方共享信息的公司、组织、个人等相关实体的名称和身份
是否有交互	指共享信息的过程中是否涉及用户与第三方的交互
与第三方共享的目的	指共享信息的目的和原因，例如推广、营销、交易等
与第三方共享的方式	指共享信息的具体方式，例如传输、出售、租赁等
与第三方共享的个人信息类型	指与第三方共享的个人信息类型，例如姓名、地址、联系方式等
用户的选择	指用户是否可以选择是否与第三方共享信息
用户选择所造成的影响	指用户选择是否与第三方共享信息所产生的影响，例如对服务的使用、优惠的享受等
第三方对数据的使用	指第三方对共享的个人信息的使用方式
第三方使用数据的目的	指第三方使用共享的个人信息的目的和原因
对于第三方的约束	指第一方对第三方使用共享信息的行为所采取的约束和措施，例如签署保密协议、限制使用范围等

3.2 数据标注方法及过程

本研究采用了CA4P-483中文隐私政策数据集中的隐私政策文本作为原始数据。该数据集包含了483篇中文隐私政策文本，综合任务需求、工作量和指标体系等多方面考虑，本研究从中抽取了130篇隐私政策文本用作标注的原始数据。

依托对于法律法规规定的系统梳理和对数据标注流程进行的深入调研，本研究形成了四千余字的详细标注手册，手册中进行了目录设置，便于标注人员翻阅。标注手册详细介绍了标注任务的基本情况，展示了整体的指标体系，并对于每一个标签类别的含义、标注规则进行了系统阐述，给出了标注示例和相关注意事项。标注手册详情参见附录。同时，本研究利用Excel制作了一个简易的标注工具，如图3-2所示，标注人员可以通过参考预设好的下拉菜单，进行方便、快捷、准确的文本标注。

在正式开始标注前，首先对于经过筛选的130篇隐私政策文本进行数据预处理。数据预处理主要包括两个部分：一是对隐私政策进行分句。本研究中的分句是根据中文标点符号，包括句号、分好、叹号、问号和省略号等，从而对隐私政策进行句

子级的标注。二是使用Python中的jieba库的停用词表，去除隐私政策文本中的停用词，通过消除文本中一些无意义或过于频繁出现的词汇，提高文本处理效率和准确性。

图3-2 标注工具示意图

在完成数据预处理后，以人工的方式进行了一次初步的数据清洗。通过对于数据集中数据进行逐条检查，删去其中一些无关的符号或少量乱码。对于按照标点符号进行分句的条件下，分句结果不准确的句子进行重新分句。这主要是由于一些隐私政策标点符号不规范导致。

在2023年3月23日至4月3日期间，由本文作者对于130篇经过数据预处理和数据清洗后的隐私政策文本进行标注。标注者本科就读于计算机专业，参与过法律数据标注平台研发的相关项目，经过前期系统梳理隐私政策相关的法律法规规定，已经具备了较好的完成隐私政策文本标注任务的能力。同时，标注手册为标注者本人制定，因此标注者已经充分学习领会了标注任务的需求。在全面投入标注工作前，进行了对于5篇中等长度隐私政策文本的试标注，统计得出标注一篇隐私政策文本平均需要25分钟左右。

在标注期间，标注者每天标注时长不超过6小时，以此保证不因为疲劳等原因影响标注质量。同时，标注者严格按照此前指定的标注手册的规则进行标注。邀请了一名从事法律相关专业的且对于隐私政策和个人信息保护相关法律较为了解的人士，对于每天的标注结果按照10%左右的比例进行抽检，抽检的方式是对于被抽到的标注结果进行逐条审查，比对是否符合标注手册的规定，如果错误率高于5%，则需要对于该篇文档进行重新标注，如果错误率高于10%，则需要整批打回。此外，每一个不

同标注日会对于两篇同样的文档进行重复标注，通过计算Cohen's Kappa系数[55]检验同一个标注者在不同天之间的标注标准是否一致，从而避免标注者因为可能存在的对于标注规则理解变化导致的标注标准不一致问题。

标注结束后，对于130篇经过多层次标注的隐私政策数据集进行系统审查和整理，全面考量标注质量是否符合要求。标注数据存储于一系列的.csv格式文档中，每一个文档存储一篇隐私政策文本内容。文档的第一列为被标注的文本内容，第二列为一级标签，第三列为二级标签。其中，由于“其他通用信息”这一一级标签没有二级子标签内容，统一将其二级标签设为“其他通用信息”，以方便后续数据处理。

最终，本研究成功形成了一个包含对于130篇隐私政策文本句子级标注的安卓应用中文隐私政策数据集。数据集被命名为“CPP-130”，即“Chinese privacy policy datasets of 130 texts”，该数据集已经在GitHub平台上进行了开源共享[56]。

3.3 数据集统计数据及对比分析

本节对于CPP-130中文隐私政策数据集的基本情况进行介绍，并将其与国外成熟通用数据集进行对比。

表3-2展示了CPP-130数据集的基本统计信息。数据集中共有17358条经过标注的文本信息，总共包含了130篇隐私政策文本，每个隐私政策文本位于一个csv格式的文件中。平均每个篇目有133.5条左右的隐私政策文本数据。数据内容按照文本顺序排列。数据集共设置了7个一级标签和38个二级标签，共标注了27642个标签。

表3-2 CPP-130数据集基本统计信息

信息名称	内容
文本总条数	17358
篇目数	130
每篇平均条数	133.5
一级标签类别数	7
二级标签类别数	38
标签总个数	27642

表3-3显示了CPP-130的一级标签分布情况。可以看出，CPP-130数据集的一级标签之间存在着一定程度的分布不均衡现象。具体来说，其他通用信息类的数量较多

达到了七千条以上，这是由于隐私政策文本中，本就三分之一以上的内容没有描述具体的数据实践，而是进行一些有关运营者等方面的基本信息阐释。第一方收集_使用、与第三方共享_转让_公开、用户访问_编辑_删除的方法和数据安全这几个类别的数量相对比较均衡，而条款更改和面向特定人群用户条款这两类的标签偏少。主要原因可能是当前隐私政策中，对于第一方收集_使用、与第三方共享_转让_公开、用户访问_编辑_删除的方法等方面的描述较为丰富，同时这三类内容本身子类别较多，占据了隐私政策的绝大部分内容。条款更改和面向特定人群用户条款内容比较简单，不需要过多阐述。同时也观察到，130篇隐私政策文件中，有43篇隐私政策文件存在着一级标签不全的现象，共缺少了72次，其中条款更改和面向特定人群条款所占比重较高。这表明，当前隐私政策不合规现象仍然较为严重，对于条款更改和面向特定人权条款缺乏的情况较为高发。通过与OPP-115数据集进行对比，发现数据不平衡的现象应当是隐私政策领域的常见现象。例如OPP-115数据集中，最多的一个类别为出现了8956次，而最少的仅出现了90次。这对于数据集最终应用于机器学习任务的效果影响并不显著。

表3-3　CPP-130数据集一级标签分布情况

标签名称	标签数量
第一方收集_使用	3317
与第三方共享_转让_公开	2196
用户访问_编辑_删除的方法	2138
数据安全	1717
条款更改	419
面向特定人群用户条款	497
其他通用信息	7074

数据集共有38个二级标签，详见表3-4所示。二级标签的分布同样不够均匀，这也是由隐私政策文本固有属性所致。例如最少的二级标签为“用户编辑、访问、删除的方法”中的“操作的数据类型”，仅有7条。而最多的则是“数据安全”中的“安全措施”，有1107条。由于部分标签数量太少，且与统一一级标签下的其余二级标签数量差距较大，在后续进行模型训练时对于此类标签进行了删去处理。在未来的数据集研发讨论中，可以进一步依据标签体系，搜集更多的此类标签的文本数据，以此来进一步提升数据集表现。由于部分标签数量过少，本数据集中对于此类二级

标签进行了剔除，最终保留了包含7个一级标签和21个二级标签的标签体系。经过统计，130篇隐私政策文档中，几乎所有隐私政策文档均存在着二级标签缺乏的现象，这也说明我国的安卓应用隐私政策合规仍然任重道远。

表3-4 CPP-130数据集二级标签分布情况

第一方收集、使用					
何时采集	749	采集的个人信息类型	466	用户选择所造成的影响	419
对数据的使用	837	使用的目的	192	是否有交互	131
采集的渠道	93	用户的选择	31	采集的目的	375
与第三方共享、转让、公开					
与第三方共享的目的	338	第三方对数据的使用	143	是否有交互	65
对于第三方的约束	355	第三方相关实体	147	用户的选择	10
与第三方共享的个人信息类型	167	第三方使用数据的目的	13	与第三方共享的方式	16
用户选择所造成的影响	49	何时与第三方共享	782		
面向特定人群用户条款					
群体类型	196	用户选择	41	供应商动作	256
数据安全					
安全措施	1107	跨境传输	154	数据存储/跨境传输的目的	18
发生安全事故的处理方式	234	数据存储期限	168		
用户编辑、访问、删除的方法					
用户可以进行的操作	1066	操作对用户造成的影响	217	供应商动作	537
操作的途径	294	操作的数据类型	7		
条款更改					
更改原因	99	用户选择	124	告知方式	143

前文已经介绍过，OPP-115是当前世界范围内最为成熟、应用最为广泛、最为权威的隐私政策英文数据集。本研究中，主要参考了OPP-115数据集的设计和研发思路，针对于安卓应用的中文隐私政策数据开展数据标注任务。相较于OPP-115数据集，本

研究中的文档数量增加了15篇，文本条数比OPP-115略少，一级类别数也略少。这是由于中英文文本的语言特性不同，我国与域外（如欧洲、美国等地）的法律法规规定也存在不同。总的来说，本数据集的标注方式和标注结果已经与OPP-115数据集相当接近，具有着发展为安卓应用中文隐私政策领域高质量标注数据集的潜力。

表3-5 数据集对比

数据集名称名称	隐私政策文档个数	文本条数	一级类别数
OPP-115	115	23194	10
CPP-130	130	17358	7

第4章 安卓应用中文隐私政策文本分类模型研究

4.1 模型设计

4.1.1 TextCNN

TextCNN是由Kim在2014年提出的一种用于文本分类的卷积神经网络模型[57]，全称是Convolutional Neural Networks for Sentence Classification，旨在解决文本分类任务中的特征提取和表示问题。

TextCNN的主要原理是利用卷积神经网络的特性来对文本进行特征提取。它使用多个不同大小的卷积核在文本上进行卷积操作，并通过池化操作提取出文本的局部特征。最后，将池化后的特征拼接在一起，输入到全连接层进行分类。

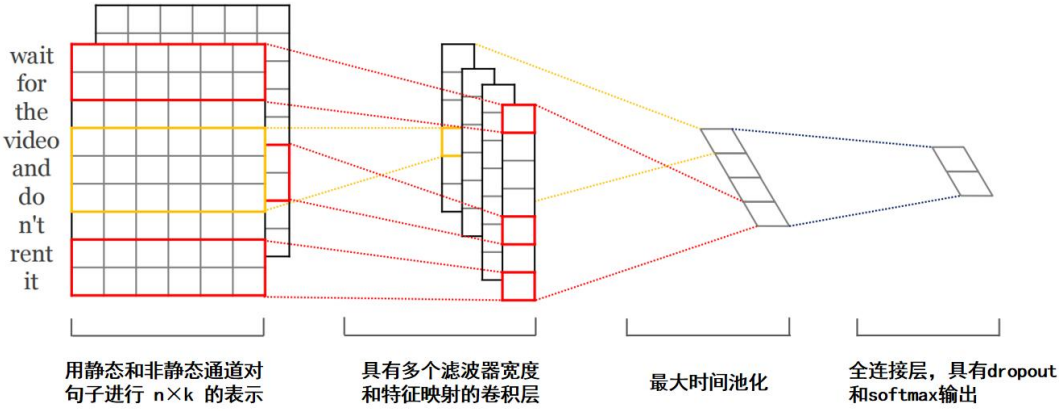

图4-1 TextCNN模型示意图

具体来说，TextCNN将输入文本表示为一个矩阵，其中每行表示一个词或字符的词向量。这个矩阵经过一系列卷积和池化操作，通过滑动不同大小的卷积核在文本上提取局部特征。卷积操作可以捕捉到不同长度的语言模式，从而有效地捕获文本中的特征。池化操作则用于提取最显著的特征，即每个特征图中的最大值。例如，对于内容为句子"I like this movie very much!"的输入，句子中共有7个单词或标点符号，对其进行基于向量化的处理，生成一个维度为5的矩阵，尺寸为[1, 7, 5]。经过向量化后的表示被馈入模型，经历了三个卷积层，其中卷积核的尺寸分别为(2, 5)，(3, 5)，(4, 5)。卷积层的输出生成了特征图，其形状分别为[1, 6]，[1, 5]，[1, 4]。接下来，应用最大池化操作对这些特征图进行处理，得到的特征图形状为[1, 2]，[1, 2]，[1, 2]。

最后，通过拼接这些池化后的特征图，形成了一个维度为[1, 6]的整体特征表示，该特征表示被用于进行两类别的分类任务。值得注意的是，上述描述中的形状尺寸中的第一个维度表示批处理大小，为了简化叙述，此处假设批处理大小为1，在实际情况中，该值可以是任意值。

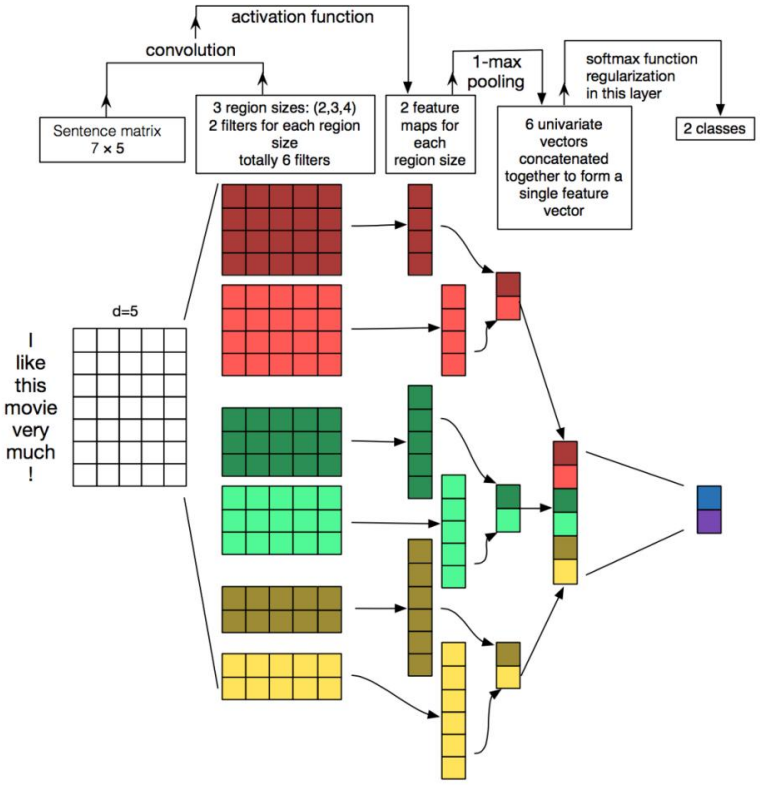

图4-2 TextCNN模型原理阐释[57]

TextCNN的主要有以下四方面优势。一是局部感知性，即通过卷积操作和池化操作，TextCNN能够有效地捕捉文本中的局部特征，不受文本长度的限制。这使得它在处理长文本和短文本时都能取得良好的效果。二是参数共享，即由于卷积操作具有参数共享的特性，TextCNN在处理文本时能够高效地学习到共享的特征表示，减少了参数数量，降低了模型复杂度。三是多尺度特征提取，即通过使用不同大小的卷积核，TextCNN可以同时捕捉不同尺度的特征，从而更好地表示文本的语义信息。四是简单有效。相对于其他复杂的模型结构，TextCNN具有简单的网络结构和训练方式，易于理解和实现，并且在很多文本分类任务上都取得了较好的性能。

总的来说，TextCNN通过卷积和池化操作，能够在文本分类任务中有效地提取和表示文本的特征，具有较好的性能和较低的复杂度，目前已经成为了一种常用的文本分类模型，对于隐私政策文本分类任务有着较大的应用潜力。

4.1.2 BERT

BERT（Bidirectional Encoder Representations from Transformers）是由Google在2018年提出的一种预训练模型，它在自然语言处理领域引起了广泛关注[58]。BERT的提出极大地推动了自然语言处理任务的性能，具有重要的研究和实际应用价值。

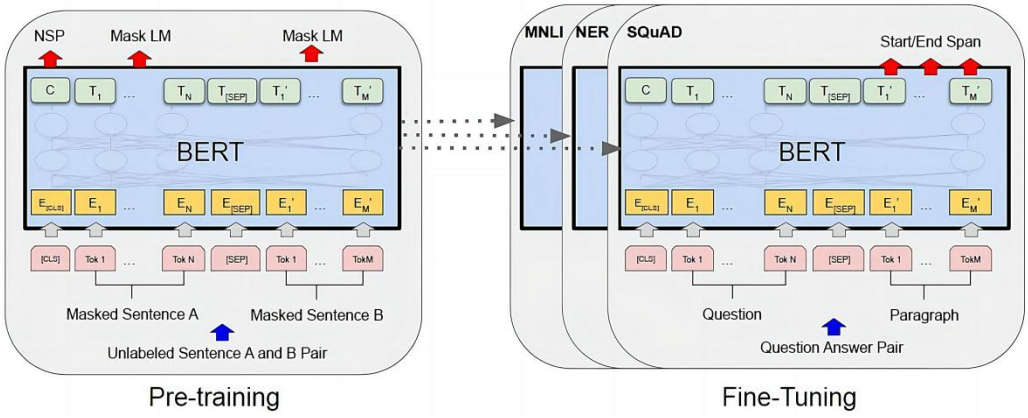

图4-3 BERT的预训练和微调过程[58]

BERT的主要原理是基于Transformer模型和无监督的预训练机制。Transformer是一种基于自注意力机制的神经网络模型，能够捕捉句子中的上下文信息。BERT通过预训练的方式，从大规模的无标注文本数据中学习到丰富的语言知识，并将这些知识应用于下游自然语言处理任务中。

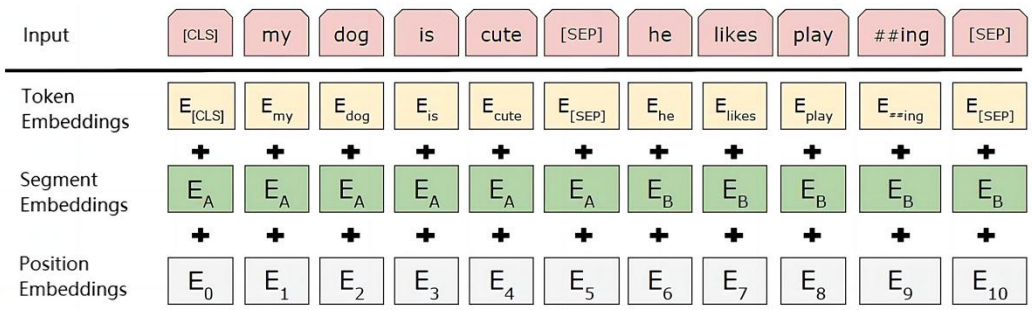

图4-4 BERT输入表示[58]

具体来说，BERT通过双向编码器（Bidirectional Encoder）来学习句子中的表示。它采用了Transformer的编码器结构，包括多个编码器层，每个编码器层由多头自注意力机制和前馈神经网络组成。在预训练阶段，BERT使用两个任务来学习句子表示：掩码语言建模和下一句预测。掩码语言建模任务要求模型根据上下文预测被掩盖的词语，从而使模型能够理解句子中的语义和语法关系。下一句预测任务要求模型判断两个句子是否是连续的，以进一步提高模型对句子关系的理解能力。

BERT的主要优势体现在以下四个方面：一是上下文感知能力，即BERT能够充分利用上下文信息，使得BERT能够更好地理解句子中的语义和语境。二是预训练与微调，即BERT采用了预训练和微调的策略，预训练阶段通过大规模无标注数据进行训练，学习通用的语言表示，微调阶段则使用有标注数据进行具体任务的训练，从而使得模型在特定任务上能够表现出色。三是上下文无关性，即可以将BERT应用于不同的自然语言处理任务，如文本分类、命名实体识别、句子关系判断等，而无需对模型进行大的改动。四是强大的表示能力，即BERT的深层网络结构和大规模预训练使其具有强大的表示能力，能够学习到丰富的语言知识和语义表示，能够处理复杂的自然语言处理任务。

总之，BERT通过采用Transformer模型和无监督预训练机制，提供了一种强大的句子表示学习方法。它的提出极大地推动了自然语言处理领域的发展，取得了显著的研究成果和实际应用效果。隐私政策文本分类是文本分类问题的一种，应用BERT能够有效地解决隐私政策文本分类问题。

4.1.3 BTC模型

BERT与TextCNN都是在文本分类任务上表现出色深度学习模型，本研究中考虑将二者进行组合，组成BTC模型（BETR-TextCNN Model），用于对于安卓应用的中文隐私政策文本进行分类。

隐私政策文本分类任务是一类特殊的文本分类任务。隐私政策文本长短不一，既有长句也有短句，很多句子需要结合上下文语境综合判断才能辨析含义。同时，隐私政策的文本结构复杂，通常包含大量的法律术语、技术术语和复杂的句子结构，又囊括各种法律意义上的概念。

通过结合BERT和TextCNN，可以充分利用BERT的上下文理解能力、词汇表覆盖范围以及TextCNN的特征提取能力，有效的针对隐私政策文本的特点完成文本分类任

务。本研究中将BERT与TextCNN组合使用的方法，是利用BERT模型每一层的模型输出的隐藏状态作为TextCNN模型的输入，随后使用TextCNN模型进行训练，最终输出训练结果，模型的建立方法如下：

首先使用预训练的BERT模型对输入的文本进行编码。BERT模型能够捕捉上下文信息和语义关系，将文本转化为高维的词向量表示。随后，提取BERT每一层的隐藏状态。从BERT模型输出中提取每一个的隐藏状态，包括embedding层和多个Transformer层的输出。这些隐藏状态代表了不同层次的抽象和语义信息。接着，将隐藏状态作为TextCNN模型的输入。将BERT每一层的隐藏状态拼接在一起，作为TextCNN模型的输入。这样做的目的是将BERT模型学到的丰富特征传递给TextCNN模型，使其能够更好地捕捉文本中的局部特征和多层特征融合。最后，使用TextCNN模型进行训练。将拼接后的隐藏状态输入到TextCNN模型中，进行文本分类的训练。TextCNN模型通过多个卷积核和池化层，能够提取局部特征和整合多层特征，进一步增强文本的表征能力和分类性能。模型结构如图4-5所示。

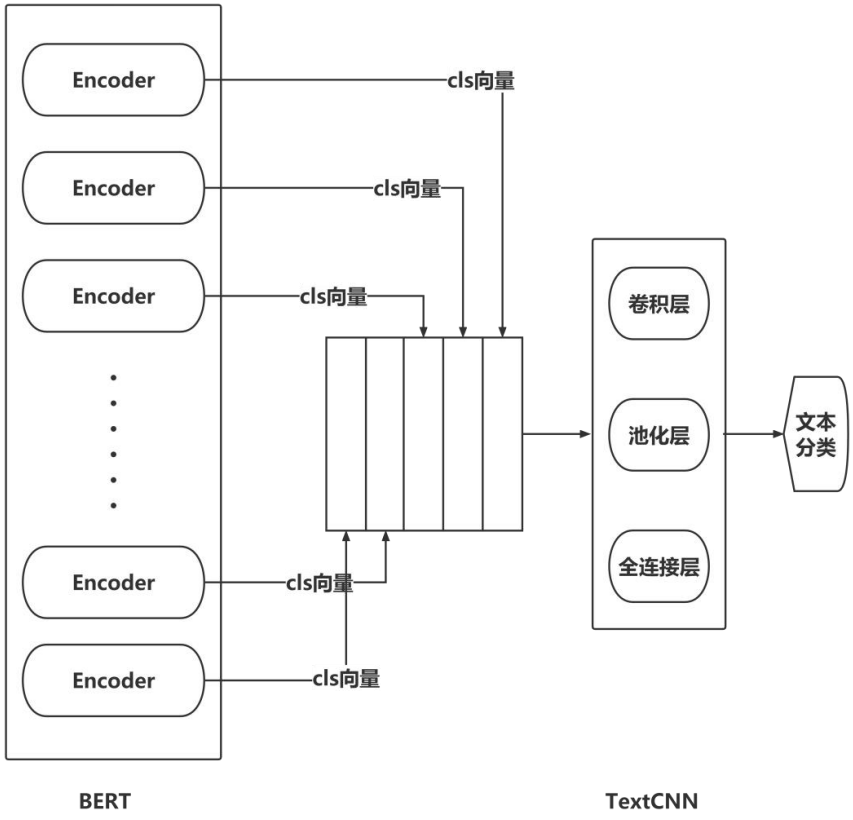

图4-5 BTC模型示意图

具体而言，首先定义一个名为BertTextModel_encode_layer的模型类。在构造函数中，初始化BERT模型和TextCNN模型，并将BERT模型加载预训练的参数。

其次，进行前向传播过程。在前向传播函数forward中，输入x包含了BERT模型的输入数据，即input_ids、attention_mask和token_type_ids。调用BERT模型的bert方法，传入输入数据，并设置output_hidden_states为真值，确保输出中包含隐藏状态的信息，得到的输出包含了BERT模型的各层隐藏状态。

再次，提取隐藏状态。从输出中中获取每一层的隐藏状态，存储在hidden_states中。其中，hidden_states的形状为[13, batch_size, seq_len, hidden]，代表了BERT模型的13个隐藏层。

接着，提取特征向量。通过遍历从第2层到第12层的隐藏状态，将每一层的第一个标记（即[CLS]标记）的隐藏向量提取出来，拼接在一起。这样得到的cls_embeddings的形状为[batch_size, 12, hidden]，代表了BERT模型每一层的特征向量。

最后，应用TextCNN模型。将得到的特征向量cls_embeddings作为输入传递给TextCNN模型，进行文本分类任务的训练。TextCNN模型会对每一层的特征向量进行卷积和池化操作，并最终输出分类结果。

```python
class BertTextModel_encode_layer(nn.Module):
    def __init__(self,hidden_size=parsers().hidden_size,class_num=parsers().class_num,filter_sizes=parsers().filter_sizes):
        super(BertTextModel_encode_layer, self).__init__()
        self.bert = BertModel.from_pretrained(parsers().bert_pred)
        for param in self.bert.parameters():
            param.requires_grad = True
        self.linear = nn.Linear(hidden_size, class_num)
        self.textCnn = TextCnnModel(hidden_size=hidden_size,class_num=class_num,filter_sizes=filter_sizes)
    def forward(self, x):
        input_ids, attention_mask, token_type_ids = x[0], x[1], x[2]
        outputs = self.bert(input_ids=input_ids, attention_mask=attention_mask,
                            token_type_ids=token_type_ids,
                            output_hidden_states=True
                            )
        hidden_states = outputs.hidden_states
        cls_embeddings = hidden_states[1][:, 0, :].unsqueeze(1)
        for i in range(2, 13):
            cls_embeddings = torch.cat((cls_embeddings, hidden_states[i][:, 0, :].unsqueeze(1)), dim=1)
        pred = self.textCnn(cls_embeddings)
        return pred
```

图4-6 BTC模型实现的关键代码

BERT以无监督的方式进行预训练，可以捕捉丰富的语义信息和上下文关系，从而能够理解隐私政策文本中的复杂语义和隐含信息。同时，BERT模型具有多层级表示，每一层都捕捉了不同抽象级别的语义特征。因此，将BERT每一层的隐藏状态作为TextCNN的输入，使得TextCNN能够同时利用多个抽象级别的特征，可以更好地捕

捉文本中的语义信息和模式，进而利用这种多层抽象特征的组合能够提供更全面、更准确的文本表示。此外，还可以将BERT中丰富的语义表示引入到TextCNN中，进而增强了模型对文本特征的理解和表达能力。

TextCNN模型具有对局部特征的敏感性，可以捕捉文本中的局部模式和特征。与BERT结合使用时，TextCNN可以从BERT的隐藏状态中提取局部特征，并与自身的卷积操作相结合，从而更全面地建模文本的结构信息和局部上下文。这种融合能够提供更具语言结构和上下文感知能力的特征表示。

通过结合BERT和TextCNN的优势，可以期望在隐私政策文本分类任务中提升模型性能。BERT提供了更准确和丰富的语义表示，而TextCNN具有对局部特征的敏感性和整体建模能力。通过组合使用它们，可以充分利用两者的优点，提高文本分类模型的准确性和泛化能力。这种组合方法能够更好地解决隐私政策文本分类任务中的挑战，提升模型对文本的理解和分类能力。

4.1.4 双层级联分类法

一次性多分类法和双层级联分类法都可以用于处理具有两层标签的隐私政策数据分类任务。

一次性多分类法是指将两层标签同时作为分类目标，通过训练一个多类别分类模型来预测所有标签的取值。该方法可以同时考虑两个标签之间的关联性，但由于需要同时预测多个类别，可能会面临标签之间的类别不平衡和样本稀疏的问题。

双层级联分类法是指将两层标签分为两个独立的分类任务进行处理。首先，训练一个模型来预测第一层标签的取值，然后根据第一层标签的预测结果，再训练一个模型来预测第二层标签的取值。这种方法可以有效地解决类别不平衡和样本稀疏的问题，并且可以充分利用第一层标签的信息来提高第二层标签的预测性能。

双层级联分类法相较于一次性多分类法具有以下优势：一是数据分布更均衡。由于双层级联分类法将两个标签分为两个独立的任务，每个任务的数据分布相对较为均衡，避免了一次性多分类中可能存在的类别不平衡问题。二是利用第一层的信息提升性能。通过利用第一层标签的预测结果作为第二层标签的输入，双层级联分类法可以充分利用上一层标签的信息，提高了第二层标签的预测性能。三是简化模型复杂度。相对于一次性多分类法，双层级联分类法可以将复杂的多类别分类任务拆分为两个独立的二分类任务，从而简化了模型的复杂度和训练的难度。

总之，双层级联分类法在处理具有两层标签的隐私政策数据分类任务时具有优势，能够克服类别不平衡和样本稀疏的问题，并且通过利用上一层标签的信息提高了分类性能。

因此，本研究中采用双层级联分类法构建隐私政策文本分类模型。训练方法是先训练一个用于对一级标签进行分类的BTC模型，再训练七个用于相应一级标签的二级子标签的BTC分类模型。使用时先将文本输入进一级标签对应的BTC模型进行分类，根据一级标签分类结果，选择对应的二级标签分类器，再将文本输入进二级标签分类器进行分类。最终输出包含两级标签的分类结果。

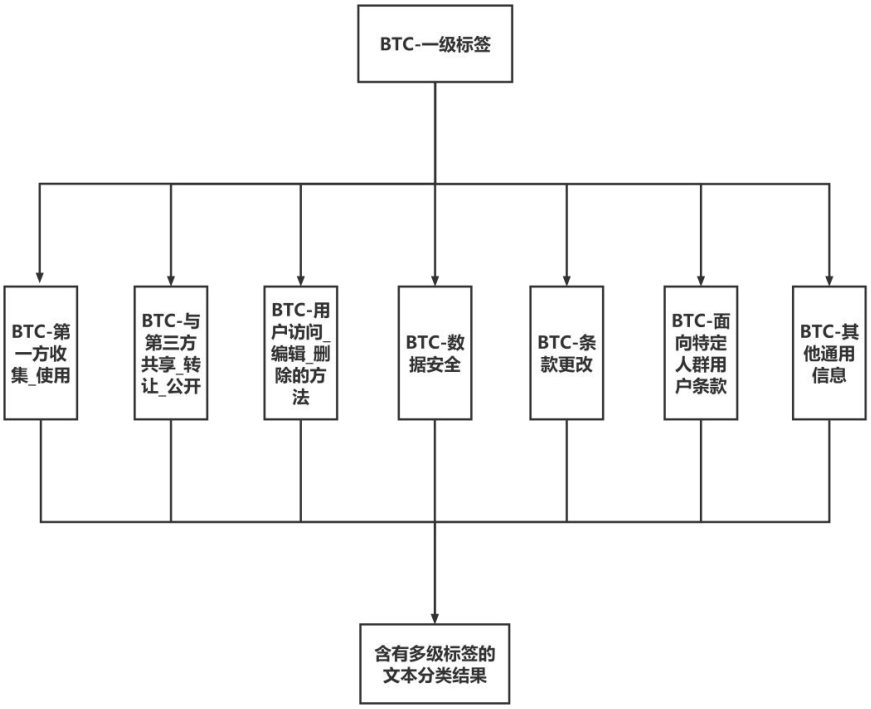

图4-7 双层级联分类法隐私政策文本分类模型示意图

4.2 模型训练与测试

4.2.1 模型训练

本节将描述模型的训练过程。

首先，从训练数据集和验证数据集中读取数据。两个数据集都分别包括了文本数据和对应的标签。使用DataLoader对数据进行批量处理。数据加载器会将数据划分

成小批量，每个批次的大小由参数batch_size决定，并进行随机打乱以增加数据的多样性。

为了优化模型的参数，本研究使用了使用了AdamW优化器。优化器会更新模型中的可训练参数，使其逐渐逼近最优解。在训练过程中，采用交叉熵损失函数(CrossEntropyLoss)作为模型的目标函数。该损失函数用于衡量模型预测结果与真实标签之间的差异，并通过最小化损失来优化模型。

随后进行多个epoch的迭代训练。在每个epoch中，首先通过train()函数进行模型的训练，该函数接受模型、设备、训练数据加载器和优化器作为输入，通过前向传播计算模型的预测结果，然后根据预测结果和真实标签计算损失，并利用反向传播和优化器更新模型的参数。训练过程结束后，使用dev()函数对模型在验证数据集上的性能进行评估。该函数接受模型、设备和验证数据集的加载器作为输入，通过前向传播计算模型的预测结果，并计算性能指标来衡量模型的准确性和泛化能力。

4.2.2 评价指标

本节将介绍用于评价深度学习模型性能的指标，包括准确率、召回率和F1值。这些指标被广泛应用于分类任务的性能评估，可以用于衡量模型的分类准确性、覆盖率和综合表现。

准确率（Precision）是评估分类模型性能的重要指标之一，它衡量了模型在预测为正例的样本中真正为正例的比例。准确率表示模型预测为正例的样本中有多少是真正的正例，它可以用于判断模型的分类能力。准确率的计算公式如下：

$$Precision = \frac{TP}{TP + FP} \tag{4-1}$$

其中，TP 表示真正例 (True Positive)，即模型正确预测为正例的样本数量；FP 表示假正例 (False Positive)，即模型错误预测为正例的样本数量。准确率的取值范围为0到1，数值越高表示模型预测为正例的样本中真正为正例的比例越高，模型的分类能力越强。

召回率（Recall）衡量了模型对正例样本的覆盖能力，即模型正确预测为正例的样本占所有真实正例样本的比例。召回率越高，模型对正例样本的识别能力越强。召回率的计算公式如下：

$$Recall = \frac{TP}{TP + FN} \tag{4-2}$$

TP 表示真正例 (True Positive)，即模型正确预测为正例的样本数量；FN 表示假反例 (False Negative)，即模型错误预测为反例的样本数量。其中，真正例是模型正确预测为正例的样本数量，假反例是模型错误预测为反例的样本数量。

F1值是综合考虑准确率和召回率的指标，用于评估模型在正例和反例样本上的综合性能。它是准确率和召回率的调和平均值。F1值的计算公式如下：

$$F1 = \frac{2 * Precision * Recall}{Precision * Recall} \quad (4\text{-}3)$$

F1值的取值范围在0和1之间，值越接近1表示模型在正例和反例样本上的综合性能越好。

通过计算准确率、召回率和F1值，可以综合评估深度学习模型在分类任务中的性能表现。这些指标可以用于比较模型的分类能力、覆盖率和综合表现，从而指导模型的改进和优化。

4.2.3 实验设置

本次实验的超参数设置如表4-1所示。

表4-1 模型超参数设置

超参数	设置情况
词向量维度	200
batch size	64
迭代次数	10
学习率	0.00001
卷积核个数	32
卷积核大小	[2,3,4]
dropout	0.2

本次实验的实验环境设置如表4-2所示。

表4-2 实验环境设置

名称	配置
操作系统	Windows 11 64位
CPU	Intel Core i9-13900H
内存	16G
GPU	NVIDIA GeForce RTX 4060 Laptop GPU
软件环境	Python3.8; torch 2.0.0+cu118

4.2.4 实验及结果分析

为了测试本文提出的BTC模型的效果，本章选取了TextCNN、Word2Vec-TextCNN、BERT与BTC模型展开对比实验。

为Word2Vec-TextCNN的效果，本研究使用了腾讯AI Lab发布的的200维中文预训练词向量作为，对文本进行了初始化[59]。该预训练词向量包括超过1200万个中文词语是在大规模高质量数据上进行预训练得到的。对于BERT模型，本研究使用的中文嵌入是预训练的bert-base-chinese模型，具有12层、768隐藏层大小、12个注意力头和1.1亿个参数。

实验使用上一章中研发的CPP-130数据集，并按照0.7:0.15:0.15的比例划分乐训练集、验证集和测试集。

先使用4种不同的模型结构进行隐私政策文本的一级标签分类模型训练，找出效果最佳的模型结构。随后，使用在隐私政策文本一级标签分类模型训练任务中表现最佳的模型结构，开展隐私政策文本二级标签分类模型训练。最终将一级分类模型与二级分类模型进行双层级联，形成多级的文本分类器。

表4-3 不同模型一级标签分类效果对比

模型名称	Precision	Recall	F1
TextCNN[57]	0.842	0.860	0.851
Word2Vec-TextCNN[59]	0.894	0.855	0.874
BERT[58]	0.887	0.868	0.878
BTC	0.928	0.932	0.930

表4-3显示了不同模型在一级标签分类任务上的效果对比。从结果来看，BTC模型在Precision、Recall和F1这三个评价指标上取得了最高的分数，分别为0.928、0.932和0.930。其次是Word2Vec-TextCNN模型和BERT模型。TextCNN模型则表现相对较弱。总的来说，BTC模型因为综合了BERT和TextCNN两种模型的优势，即BERT模型具有的强大的语义理解能力和上下文感知能力，以及TextCNN模型具有的较好的特征提取能力。BERT模型和Word2Vec-TextCNN模型的表现也较好，这可能是因为这两个模型使用了预训练的思想，通过在大量语料上进行预训练，可以有效地提升模型性能。

表4-4 BTC模型一级标签分类效果对比

标签名称	Precision	Recall	F1
第一方收集_使用	0.921	0.921	0.921
与第三方共享_转让_公开	0.937	0.881	0.901
用户访问_编辑_删除的方法	0.918	0.912	0.915
数据安全	0.953	0.961	0.957
条款更改	0.885	0.871	0.878
面向特定人群用户条款	0.974	0.998	0.987
其他通用信息	0.936	0.950	0.943

表4-4展示了使用BTC模型进行隐私政策文本一级标签分类的效果，列出并对比了七个一级标签的预测结果。总的来说，BTC模型对于七个一级标签均有着较好的预测能力，在一级标签分类任务上表现均衡且出色。分类效果最好的是“面向特定人群用户条款”，这一类别的文本虽然偏少，但是特征尤为显著，几乎所有文本都会提及类似于“未成年人”、“老年人”、“监护人”等类似的表述，即涉及到人群类别相关的表达，因此模型对其分类效果出色。相对而言，分类效果最差的是“条款更改”这一标签，分类效果差的可能原因是“条款更改”的文本数量最少，并且这一类别的文本特点不突出、句子长度差异大，导致模型难以学习到相应的特征。

由此，实验验证了BTC模型对隐私政策文本一级标签具有着最好的分类能力。在二级分类模型训练时，需要进行分类的隐私政策文本是完全相同的，只是标签信息不同。因此可以参照一级标签分类结果，使用BTC模型开展二级分类模型训练。

表4-5 BTC模型二级分类模型效果

模型名称	Precision	Recall	F1
BTC-第一方收集_使用	0.887	0.868	0.878
BTC-与第三方共享_转让_公开	0.897	0.789	0.839
BTC-用户访问_编辑_删除的方法	0.891	0.946	0.918
BTC-数据安全	0.945	0.981	0.962
BTC-条款更改	0.917	0.922	0.919
BTC-面向特定人群用户条款	0.893	0.887	0.890

表4-5展示了使用BTC模型进行二级隐私政策文本标签分类的效果。可以看出，BTC模型应用于各个二级标签分类任务时表现都较为出色。其中表现最好的是“数

据安全”，原因可能在于其子标签个数相对较少，同时文本总数适中、数据平衡性较好，子类之间表述区别较显著。分类表现最差的是“与第三方共享_转让_公开”这一类别，主要原因可能是数据平衡性较差，同时第三方的相关表述可能较为模糊。

综上所述，本章的理论分析和实验结果验证了BTC模型在隐私政策数据分类任务中的优越性。在进行二级标签分类模型训练时，根据一级标签分类结果选择BTC模型作为训练模型，以获得更好的分类效果，由此基于双层级联分类法和BTC模型结构训练出了一个用于隐私政策文本分类问题的模型。该模型可以为后续章节的工作提供支撑，更可以广泛应用于其他与隐私政策文本分类相关的问题。

第5章 安卓应用隐私合规性检测系统的设计与实现

为了将文本分类结果更好的进行可视化，同时实现隐私政策的量化评估、标签统计等功能，辅助执法机构、企业和产品用户更好的对于安卓应用隐私政策的合规性进行检测，本章将第三章提出的隐私政策评价指标体系、第四章研发的隐私政策文本分类BTC模型，应用于安卓应用隐私合规性检测系统的设计与实现过程。前端系统使用Vue进行构建，用于获取用户需要检测的隐私政策文本，并展示隐私合规性检测的可视化结果。后端系统基于Python搭建，接收前端传回的数据，输入进模型中进行处理，依据分类结果进行分析，并返回相应的检测结果。

5.1 功能分析

为了对安卓应用的隐私合规性进行检测，本文以隐私政策为切入点，设计了一种检测系统。系统从前端接收待检测文本，输入进后端模型并做相关的分析，返回包括量化评估结果、标签统计结果和标签可视化结果在内的三类检测结果。

本研究对于三位安卓隐私政策合规领域的相关人士进行了访谈，访谈问题列表详见附录。根据访谈结果，本研究总结出以下几方面需求：

一是文件批量上传。对于隐私政策合规领域的执法机关而言，一次可能需要检测成千上万的安卓应用隐私政策，因此如果逐个手工录入隐私政策文本、逐个上传、逐个检测的话，需要消耗大量时间精力在重复劳动上，因此很有必要能够批量上传隐私政策文本，并且成功上传一个文件后，后端系统就开始对其进行自动分析。对于企业法务的隐私政策合规相关负责人而言，大型企业旗下往往同时有多个平台和应用，因此也需要对于这些应用进行批量检测。同时，还需要能够根据文件名等信息快速选择想要查看的隐私政策文件的检测结果。

二是隐私政策自动化检测。隐私政策文本的复杂性和多样性使得准确地理解和解释这些文本变得具有挑战性。大量的法律术语和技术术语需要人们具备专业知识才能正确理解其含义。快速变化的法律法规和行业标准增加了合规检测的复杂性。应用开发者需要及时调整和更新其隐私政策，以适应法规的变化，这给检测工作带来了一定的困难。此外，大规模的应用数量和不断增长的用户数据也对检测工作提出了挑战，需要高效的工具和自动化检测方法来应对。最为直观的方式是通过建立量化评估体系，以打分的形式评判隐私政策的合规性。

三是隐私政策文本分类结果可视化。自动化检测往往无法完全替代人工检测，更多的是为人工检测提供一个初步参考，让检测人员仅仅需要进行审核的工作，以此来提升工作效率。因此对于隐私政策文本分类结果进行可视化是十分重要的。一方面，隐私政策文本标签类别繁多，同时可能分布于文本的各个位置。检测时检测人员实际上是希望能够以标签为核心进行检测，而不是完全依照文本顺序。另一方面，了解各个标签的数量十分重要，通过数量就可以直观看出隐私政策中对于哪些法律法规的贴合度较高，哪些则相对较低。

针对上述需求分析，本研究提出了系统的主要功能：

一是文件处理功能。该功能允许用户通过系统界面上传存储了隐私政策文本的txt文件。用户可以选择上传单个文件或多个文件。 系统采用上传即检测的模式，用户无需等待，即可立即获取检测结果，提高了效率和用户体验。此外，用户可以从已上传的文件列表中选择一个文件，以查看该文件的检测结果。文件选择功能允许用户从已上传的文件列表中选择特定文件，提供根据文件名检索的方式，用以查看该文件的检测结果。

二是量化评估功能。该功能用于对隐私政策进行量化评估，包括评估其真实性、完整性和内容集中程度。系统基于预定义的评估指标和算法，对隐私政策进行自动评估，并向用户提供评估结果，帮助用户了解隐私政策的质量和合规程度。

三是标签统计功能。该功能用于统计隐私政策中各个一级和二级标签的数量。系统自动识别和统计文本中的标签，并以直观的方式展示各个标签的数量信息，让用户能够快速了解隐私政策中各个标签的分布情况。

四是标签可视化功能。该功能通过使用不同字体颜色标记被标为不同一级标签的句子，实现标签的可视化。当用户将鼠标悬停在某个句子上时，系统会显示该句子的二级标签，以提供更详细的标签信息。这样的可视化功能能够帮助用户更直观地理解隐私政策文本中的标签分布情况，快速识别出与自己关注的标签相关的内容。

5.2 系统架构

5.2.1 前端架构

为了实现隐私政策标签可视化和文件上传功能，本研究使用了一个前端框架。该前端框架采用了基于Vue.js和Ant Design的技术栈。

在前端框架中，本研究通过使用Ant Design提供的组件库来构建用户界面。首先，创建一个顶部布局（a-layout-header），其中包含文件选择和文件上传功能。文件选择功能通过下拉列表（a-select）实现，用户可以从已上传的文件中选择查看结果。文件上传功能通过文件上传按钮（a-upload）实现，用户可以上传一个或多个存储了隐私政策文本的txt文件。

接着，创建一个侧边栏布局（a-layout-sider），用于显示隐私政策标签的统计信息和标签导航菜单。侧边栏中的菜单使用了Ant Design的菜单组件（a-menu），其中包含了不同的一级和二级标签，用于对隐私政策进行分类和导航。用户可以点击菜单项来查看对应标签的详细信息。

在主要内容区域（a-layout-content），系统展示了隐私政策文本的标签可视化结果。本研究使用了Vue.js的v-for指令来遍历文本中的句子，并根据其标签类型为每个句子应用相应的样式。当用户将鼠标悬停在某个句子上时，会显示该句子的二级标签。

前端框架整体基于Vue.js的响应式设计，可以根据用户的操作和数据的变化自动更新界面。通过使用Ant Design的组件库和Vue.js的响应式特性，本研究实现了一个直观、易用且具有良好用户体验的前端界面，为用户提供了方便的文件上传、选择和隐私政策标签查看的功能。

总的来说，本研究的前端框架主要用以支持以下功能：

一是文件处理功能。允许用户上传一个或多个存储了隐私政策文本的txt文件，允许用户从已上传的文件中选择查看检测结果。

二是隐私政策标签统计功能。统计并显示隐私政策各个一级和二级标签的数量。

三是隐私政策标签可视化功能。通过不同的字体颜色标记文本中的句子，并显示其对应的一级和二级标签。

四是量化评估结果展示功能：展示隐私政策文本的量化评估得分情况。

5.2.2后端架构

本研究使用了Python的Web框架Flask作为后端开发工具。Flask是一个轻量级的Web应用框架，提供了简单而灵活的方式来构建服务器端应用程序。

首先，创建一个Flask应用对象。同时，为了解决跨域问题，引入了flask_cors模块，并使用CORS(app)将跨域请求设置为允许。

为处理文件上传请求，定义upload()函数，并将其与/uploadFiles路由绑定。该函数通过request.files.getlist('file[]')获取上传的文件列表，然后遍历每个文件，将其保存到本地的uploads目录下。最后，将上传成功的信息以JSON格式返回给客户端。

为获取文件处理结果，定义getfile()函数，并将其与/getFileResults路由绑定。在该函数中，首先检查全局标志flag的值，以确保文件处理结果只被获取一次。然后，遍历uploads目录下的每个文件，读取文件内容，并调用eval_pol()函数对文本进行处理和分析。最后，将处理结果存储在result字典中，并将其作为JSON数据返回给客户端。

为解决跨域问题，定义after_request()函数，并使用app.after_request(after_request)将其设为请求后的处理函数。在该函数中，将响应头中的Access-Control-Allow-Origin设置为'*'，从而允许所有来源的跨域请求。

5.3 文件处理功能

文件处理功能是系统的基本功能之一，包括文件上传和文件选择两个子功能，本节将介绍这两个子功能的设计和实现方法。

5.3.1 文件上传功能

文件上传功能允许用户上传一个或多个存储了隐私政策文本的txt文件，并采用上传即检测的模式。该功能的设计目标是为用户提供一个简单、高效的方式来将隐私政策文本导入系统进行检测，功能具体如下。首先，系统提供一个用户界面，用于选择要上传的文件。用户可以通过单个文件选择或多个文件选择的方式来上传文件。其次，在文件上传过程中，系统会对文件进行格式验证，确保只接受txt文件格式。对于不符合要求的文件格式，系统会给予相应的提示信息，引导用户重新选择正确的文件。系统使用后端处理逻辑来接收上传的文件。再次，在上传完成后，系统将文件保存到指定的存储位置，以供后续的检测和处理。最后，一旦文件上传完成，系统将立即对上传的文件进行检测，以识别其中的隐私政策文本并提取相关信息。该即时检测的方式可以提高用户的操作效率和用户体验。

5.3.2 文件选择功能

文件选择功能允许用户选择查看哪一个文件的检测结果。该功能的设计目标是为用户提供一个方便的方式来查看已上传文件的检测结果，功能具体如下：

文件列表展示：系统在用户界面上展示已上传文件的列表，以便用户选择要查看的文件。文件列表应包含文件名、上传时间等相关信息。

文件选择：用户可以通过点击文件列表中的某一文件来选择查看该文件的检测结果。系统会相应地加载该文件的检测结果并在界面上展示。

检测结果展示：系统会将文件的检测结果以易于理解的方式展示给用户。展示方式可以采用表格、图表或其他形式，以清晰地呈现隐私政策文本的检测结果。

文件处理功能的设计和实现需要综合考虑用户的需求和系统的性能要求。通过提供方便的文件上传和选择方式，系统可以提高用户的使用便捷性，并有效支持隐私政策文本的检测和分析工作。

5.4 隐私政策量化评估功能

5.4.1评估指标体系回顾

在第三章中，为了建立科学、可靠的移动应用隐私政策合规评估体系，本研究主要参考《规范》和其他相关法律法规，建立了一个包括7个一级指标和38个二级指标的指标体系。这些指标涵盖了安卓应用在数据收集、共享、处理等方面的主要实践行为类型。

本节将基于这个指标体系，设计隐私政策合规性的量化评估方法，用于对移动应用隐私政策合规性进行评价。具体而言，评估方法首先将移动应用隐私政策中涉及到的一级指标和二级指标抽象为评价项，然后通过分析隐私政策中涉及到的关键词和语句，运用自然语言处理技术，对评价项进行自动化提取，最终对隐私政策的真实性、完整性和内容集中程度进行评分，即从各个角度评价安卓应用隐私政策合规性。通过这套量化评估，可以提高安卓应用隐私政策的合规性和透明度，保护用户个人隐私权益，同时为安卓应用运营者提供科学、可靠的指导和支持。

5.4.2真实性评估

当前判断隐私政策真实性有两种主要角度，一种是判断应用开发者给出的隐私政策链接是否真实，另一种则是判断隐私政策文本内容是否真实。前者是对于链接进行检测，例如测试链接是否可以正常打开，打开后是否存在内容；后者则是假定隐私政策通过了前者后，对于隐私政策的内容进行判断，判断一篇隐私政策是否真的在描述用户数据实践相关的内容。

本研究统计了经过标注的130篇隐私政策的数据，发现其中“其他通用信息”所占比例约为40%，而除了“其他通用信息”外的其余几类标签均属于数据实践相关内容。随后以40%为基准，进行人工测试，通过人工统计并判断50篇隐私政策的真实性情况与数据实践相关的内容所占比例的关系，得出其所占比例应当高于55%，这与前人研究中得出的结论基本一致[18]。因此有真实性得分R为：

$$R = \frac{\sum_{i=1}^{6} C_i}{N} \tag{5-1}$$

其中R 表示隐私政策的真实性性得分；Ci表示隐私政策中第i个类别的一级标签数量之和；N 表示隐私政策的总文本条数。R的值为一篇隐私政策中前六类标签数量的总和除以文本总条数，即为完整性得分。当R低于0.55时，则判定该隐私政策真实性不足，即为虚假隐私政策。

5.4.3 完整性评估

隐私政策完整性即是指隐私政策是否能较好的覆盖相关法律法规规定的要求，是否含有相应的内容用以告知用户。本研究依托第三章中建立的隐私政策评价指标体系，设计了隐私政策的完整性评估公式：

$$W = N \cdot \sum^{n} w_i c_i \tag{5-2}$$

其中，W为隐私政策的完整性评分，其值位于0到1之间。$N = 1/\sum_{i=1}^{n} w_i$是一个归一化因子，用于将结果缩放到[0, 1]的范围内。n表示隐私类别的数量，包括7个一级标签和21个二级标签，总计38个标签。w_i表示用户分配给第i个类别的权重，取值范围为[0, 1]。c_i表示覆盖水平，其取值为1表示已覆盖，取值为0表示未覆盖。在本研究中，经过多次测试，将一级标签的权重统一设为0.7，将二级标签的权重统一设为0.3。

5.4.4 内容集中程度评估

判断隐私政策合规性的难度，不仅在于隐私政策背后的法律法规纷繁复杂、隐私政策数量众多，还在于隐私政策本身可读性较差、内容质量不高，而隐私政策的内容集中程度，是反应隐私政策可读性和内容质量的一个重要视角。通过对隐私政策的深入阅读研究，可以发现，在隐私政策文本中，某个类别的条款描述可能在多个位置重复出现，而不同类别的描述也可能在某些位置交叉提及。这种情况下，人

们很难理清隐私政策中具体表达的内容。基于这样的思考，本研究设计了隐私政策的内容集中程度评估公式：

$$E_i = \frac{n_i - 1 - \sum_{i=1}^{n_i-1} d_{i,i+1}}{n_i - 1} \tag{5-3}$$

$$E = \frac{\arctan(\frac{\sum_{i=1}^{n_i-1} E_i}{t-1})}{\pi} \tag{5-4}$$

E_i指某一个类别的内容集中程度，n_i指某一个类别的语句数量，$d_{i,i+1}$指某个类别的两个语句中所间隔的语句数量。E指隐私政策整体的内容集中程度评分，T指隐私政策文本的总类别数。使用arctan函数的目的是对于结果进行归一化。需要特别注意评估内容集中程度时，不考虑“其他通用信息”这一类别，因为其他通用信息与数据实践无关，不应当成为内容集中程度关注的信息。

5.5 标签分类结果可视化功能

为实现标签分类结果可视化功能，本研究使用 Vue.js 框架和 Ant Design Vue 组件库。首先引入了所需的组件和图标资源，然后在 <template> 标签中定义了页面的布局结构，并将相关的组件放置在合适的位置。在 <script> 标签中，通过使用了 Vue.js 的 setup() 函数进行初始化，并定义与功能相关的逻辑和方法。

5.5.1 一级标签可视化展示

本研究通过将句子标记与一级标签关联，使用不同的字体颜色来标记被标为不同一级标签的句子。当用户点击标签列表中的标签名称时，隐私政策文本中相应的内容就会改变字体颜色。用户可以只查看一种类别的标签，也可以同时查看多种类别的标签。通过这种方式，可以直观地将句子与其对应的一级标签进行关联。定义并使用一个包含多个句子和相应标签的数据结构，通过遍历这个数据结构并根据每个句子的一级标签选择合适的字体颜色，将不同一级标签的句子标记出来，使用 v-for 指令来遍历句子列表，并通过动态绑定 CSS 类名的方式，为每个句子应用适当的字体颜色。通过为每个句子应用适当的类名，最终实现了句子标记与一级标签关联的效果。

5.5.2 二级标签可视化展示

本研究通过鼠标悬停显示二级标签来实现二级标签的可视化展示。本功能使得当鼠标悬停在某个句子上时，会显示该句子的二级标签。本研究利用 <a-tooltip> 组件来实现，通过将 <a-tooltip> 包裹在每个句子所在的 元素周围，可为每个句子添加一个悬停提示。通过使用插槽（slot）机制，将每个句子的二级标签作为提示的内容。当用户将鼠标悬停在某个句子上时，会触发提示的显示，从而展示该句子的二级标签。为句子的 元素添加鼠标悬停事件的监听器。当鼠标悬停事件被触发时，通过修改相应句子的状态来控制 <a-tooltip> 的显示与隐藏。通过 v-show 指令来动态控制 <a-tooltip> 组件的显示状态，并将当前句子的二级标签作为插槽的内容传递给 <a-tooltip> 组件。

5.5.3 标签数量统计展示

通过从后端接收返回的标签数量列表，在标签名称旁边显示该标签的数量。颜色与一级标签可视化功能中，对应的一级标签文本的颜色一致，以此来方便用户查看相应的标签数量，并与一级标签类别建立对应关系。

5.6 系统主要功能测试

在系统研发完成后，对于系统的主要功能进行测试，测试情况展示如下。

图5-1展示了系统的初始界面，可以看到，初始界面中包括了左侧的标签列表，列表可以展开查看二级标签。右上角为文件上传功能，左上角为文件选择功能。页面右下的大部分区域暂时处于空白状态。

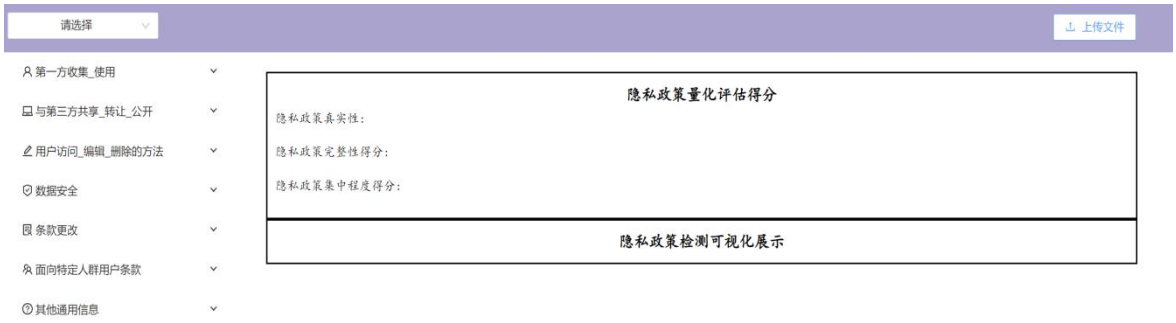

图5-1 系统初始界面

图5-2展示了系统的文件上传功能。点击上传文件后，可以从计算机的任意一个文件夹选择txt格式的文件进行上传。上传后系统将会自动检测，如果文件格式为非txt将会报错，如果文件格式符合，将会自动开始进行检测。

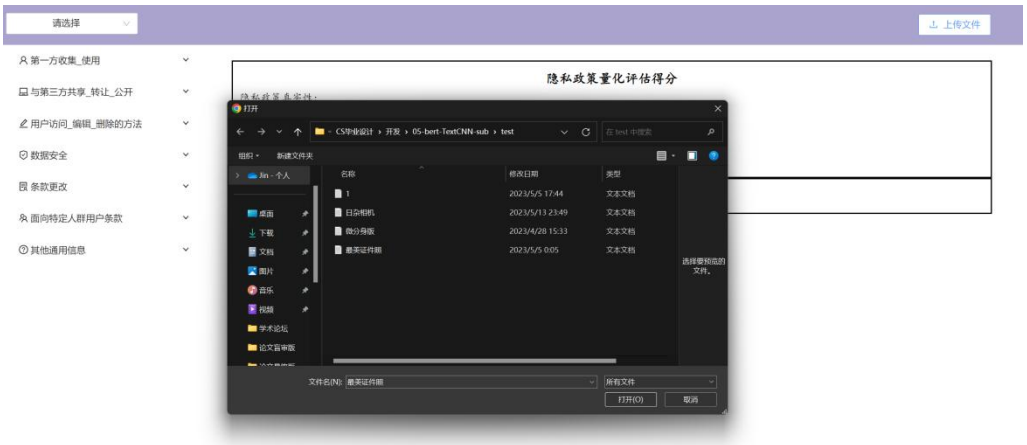

图5-2 文件上传功能

图5-3和图5-4展示了系统的文件选择和文件搜索功能，可以单击下拉框选择所有已经上传的文件，也可以根据文件名对于文件进行检索。

图5-3 文件选择功能

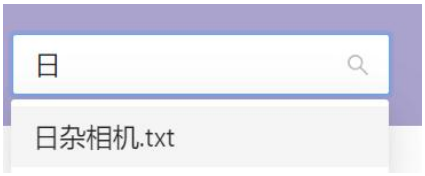

图5-4 文件搜索功能

图5-5展示了系统的标签数据统计功能。每个标签右侧的数字为这一类标签的个数。

图5-5 标签数据统计功能

图5-6展示了选择一个文件后的初始效果，系统将会显示文件的名称、各个标签的数量、三类量化评估情况，并在下方展示隐私政策的内容。

图5-6 隐私政策检测初始效果

图5-7展示了系统后端运行情况，可以看出，系统后端成功收到了指令并运行了模型。

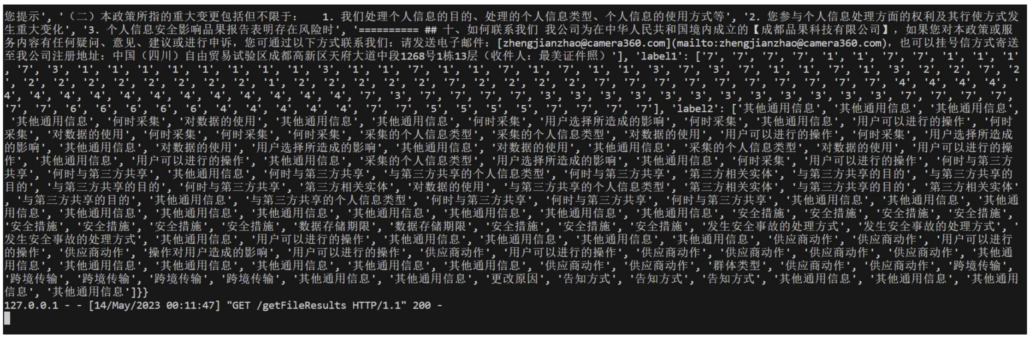

图5-7 后端运行成功

图5-8展示了单独查看一个以及标签类别的情况，可以看出，用户当前在查看“第一方收集_使用”这一标签，而下方的文本框中，相应的文字变为了红色。

图5-8 单独查看一个一级标签类别

图5-9展示了同时查看多个一级类别的情况，当前用户查看了所有类别。其中粉色字体颜色较浅，用于表示无关的“其他通用信息”，红色字体仍然表示“第一方收集_使用”，而橘黄色字体表示“与第三方共享_转让_公开”这一标签类别。

图5-9 同时查看多个一级标签类别

图5-10展示了查看二级标签类别的情况，此时用户将鼠标放置于“1.3收集、使用个人信息目的变更……”，则显示出了相应的二级标签“更改原因”。

图5-10 查看二级标签类别

总的来说，通过进行系统主要功能测试，较好的展示和检验了系统应用于安卓应用隐私合规性检测的能力及潜力，成功验证了包括文件处理、量化评估和可视化展示等功能的实现情况。

第6章 总结与展望

6.1 研究总结

在国家大力推动科技社会治理，隐私保护相关法律法规接连出台，安卓应用隐私合规性检测困难重重的背景之下，本研究以安卓应用隐私政策为切入点，通过文献回顾、交流访谈、数据标注、模型训练、系统研发等工作内容，形成了安卓应用隐私合规检测系统，具体如下：

首先，本研究建立了一套全面的隐私政策合规性评价指标。通过广泛阅读相关文献并深入研究隐私政策领域的法律法规规定，本研究系统梳理了关键要素，并提出了包含7个一级指标和21个二级指标的隐私政策评价方法。这些指标覆盖了隐私政策的核心要求，能够全面评估隐私政策的合规性，确保隐私保护的有效实施。此外，本研究还针对真实性评估、完整性得分和内容分布得分等关键方面，提供了三类量化评估手段，为隐私政策合规性评价提供了有效工具。

其次，本研究构建了高质量标注的中文隐私政策数据集。基于建立的隐私政策合规性评价指标，制定了严格的中文隐私政策文本标注手册，并对130篇隐私政策中的一万七千余条文本进行了多层级标注。通过数据清洗和标注过程，得到了一份高质量的中文隐私政策文本数据集。这一数据集的建立为后续的隐私政策分析、模型训练和算法优化提供了可靠的基础。

再次，本研究研发了更高效的隐私政策文本分类模型。采用了近年来在文本分类任务上取得显著成果的深度学习模型，如Text-CNN、BERT等，通过大量实验和分析，成功地开发出了一种高效的中文隐私政策文本分类模型BTC模型。该模型能够准确地对隐私政策进行分类和归类，为隐私政策的自动处理和分析提供了有力支持。

最后，本研究还开发了一款安卓应用隐私合规检测系统。该系统结合了建立的隐私政策合规性评价指标和文本分类模型，能够自动化地检测安卓应用的隐私政策合规性。通过该系统，开发者和用户能够方便、快速地评估应用的隐私政策合规性，提升用户对隐私保护的信任。

综上所述，本研究在隐私政策合规性评价、中文隐私政策数据集构建、文本分类模型研发以及应用实践等方面取得了重要的研究成果。这些成果不仅推动了隐私

政策研究的发展，还为隐私保护提供了实用的方法和工具，为构建可信赖的数字环境做出了贡献。

6.2 研究局限性

本研究主要具有四方面局限性：

一是合规检测范围的局限性。本研究所建立的隐私政策合规性评价指标体系覆盖了广泛的要素和指标，但仍然可能存在某些特定情境或具体领域的合规要求未完全考虑到的情况。此外，法律法规规定仍在不断更新和出台，要持续对指标体系进行维护和完善，并进一步细化和扩展指标体系，以满足更多场景的合规性要求。

二是数据集建设的局限性。尽管本研究构建了高质量标注的中文隐私政策数据集，但数据集的规模和覆盖范围仍受限制。隐私政策文本的标注过程可能存在主观性和人为偏差，且数据集中的隐私政策样本可能无法覆盖所有应用领域和行业。进一步扩充和丰富数据集，提高数据的多样性和代表性，将是未来研究的方向。同时，当前数据集只有两层标签，对于数据实践的具体内容仍然不能表示。

三是模型设计的局限性。本研究对于BERT和TextCNN模型进行组合，用于隐私政策文本分类，但不同模型的选择和参数设置可能会对分类结果产生影响。此外，模型的泛化能力和鲁棒性也需要进一步验证和改进。未来的研究可以探索更多的模型结构和算法策略，以提高分类模型的性能和适用性。

四是系统研发的局限性。虽然本研究开发了安卓应用隐私合规检测系统，但在实际应用过程中，仍然可能存在一些技术实现和性能方面的限制。系统仍然处于比较初级的阶段，在可扩展性、易用性和实时性等方面仍需要进一步优化和改进。同时，系统易用性暂时不高，仍需进一步扩展功能点，提升用户界面友好性。

6.3 研究展望

本文对未来领域内研究提出以下三点展望：

第一，要进一步建设高质量的中文隐私政策数据集。虽然本研究已经构建了一份高质量的中文隐私政策数据集，但在数据集的规模、覆盖范围和多样性方面仍有进一步的提升空间。当前中文领域的隐私政策数据集尤为缺乏，而隐私政策的语言和表达方式与不同语种和文化背景相关，因此建设中文隐私政策数据集具有重要意

义。未来的研究可以通过更广泛的数据收集和多方参与的标注过程，构建更大规模、更全面且具有代表性的数据集，以支持更深入的隐私政策研究和开发。

第二，要丰富并深入隐私政策文本分类模型的设计。本文采用了深度学习模型进行隐私政策文本分类，但仍可以进一步探索和优化模型的结构和算法。未来的研究可以尝试引入更多的自然语言处理技术、迁移学习和模型融合方法，以提升模型的分类性能和泛化能力，还可以深入研究隐私政策文本的特征提取和表示方法，以更好地捕捉文本中的隐私相关信息。此外，还可以结合以ChatGPT为代表的大模型开展研究，探索如何将其应用于隐私政策文本分类甚至隐私政策合规性检测任务，通过预训练和微调的方式，提高分类模型的准确性和泛化能力，进一步提升隐私政策合规性评价的效果。

第三，要设计用户友好的隐私合规检测系统。随着隐私保护意识的增强，用户对隐私政策的关注度也在提高。因此，未来的研究应当致力于设计用户友好的隐私合规检测系统，使普通用户能够轻松理解和评估应用程序的隐私政策合规性。这包括开发直观易懂的可视化工具、提供个性化的隐私偏好设置以及利用自然语言处理技术提供用户友好的隐私政策摘要和解释等。同时，要进一步调研需求，对于隐私政策通过提高用户对隐私政策的可理解性和可操作性，可以增强用户对个人数据隐私的保护意识和行动能力，同时也可以提升隐私政策合规检测的效率和质量。

总之，未来的研究可以继续在建设高质量数据集、丰富分类模型设计以及设计用户友好的隐私合规检测系统等方面进行探索和创新，以推动隐私政策领域的发展，并为实际应用提供更全面、准确和用户友好的支持。

结　论

在国家大力推动科技社会治理，隐私保护相关法律法规接连出台，安卓应用隐私合规性检测困难重重的背景之下，本研究以安卓应用隐私政策为切入点，通过文献回顾、交流访谈、数据标注、模型训练、系统研发等工作内容，形成了安卓应用隐私合规检测系统，具体如下：

首先，本研究建立了一套全面的隐私政策合规性评价指标，7个一级指标和21个二级指标，依托指标体系提出了隐私政策合规性的量化评估方法。其次，本研究构建了高质量标注的中文隐私政策数据集，制定严格的中文隐私政策文本标注手册，并对130篇隐私政策中的一万七千余条文本进行了多层级标注和数据清洗。再次，本研究研发了高效的隐私政策文本分类模型，通过组合BERT和TextCNN模型结果，经过大量实验和分析，成功地开发出了一种高效的中文隐私政策文本分类模型BTC模型。最后，本研究开发了一款安卓应用隐私合规检测系统。该系统结合了建立的隐私政策合规性评价指标和文本分类模型，能够自动化地检测安卓应用的隐私政策合规性。

本研究的局限性如下：一是合规检测范围有限，无法完全时效性和覆盖率；二是数据集规模和覆盖范围有限，标注过程可能存在主观性和偏差；三是模型设计需进一步改进，选择和参数设置可能影响结果；四是系统研发仍处于初级阶段，需要优化可扩展性、易用性和实时性。

未来的研究可以继续在建设高质量数据集、丰富分类模型设计以及设计用户友好的隐私合规检测系统等方面进行探索和创新，以推动隐私政策领域的发展，并为实际应用提供更全面、便捷、准确的支持。

总的来说，本研究立足于科技社会治理和个人隐私保护的背景下，通过建立隐私政策合规评价指标体系，构建高质量中文安卓隐私政策数据集，探索隐私政策文本分类模型，研发隐私政策合规性检测系统，有效利用科学技术解决社会治理问题，为增强我国个人隐私保护领域执法、监督能力做出了贡献，对提升用户的个人信息安全水平和信任感具有重要意义，是对中文隐私政策合规性检测领域的一次有价值的探索。

参考文献

[1] 欧洲联盟. 通用数据保护条例[EB/OL]. (2016)[2023-05-08]. https://gdpr-info.eu.

[2] 全国信息安全标准化技术委员会秘书处. 信息安全技术个人信息安全规范[EB/OL]. (2020-09-18)[2023-05-08]. https://www.tc260.org.cn/front/postDetail.html?id=20200918200432.

[3] 全国信息安全标准化技术委员会秘书处. 网络安全标准实践指南—移动互联网应用程序（App）使用软件开发工具包（SDK）安全指引[EB/OL]. (2020-11-26)[2023-05-08]. https://www.tc260.org.cn/front/postDetail.html?id=20201126161240.

[4] 中华人民共和国国家互联网信息办公室. App违法违规收集使用个人信息行为认定方法[EB/OL]. (2019-12-27)[2023-05-08]. http://www.cac.gov.cn/2019-12/27/c_1578986455686625.htm.

[5] 王晓宁. 移动社交APP隐私政策的合规性研究——基于20例隐私政策文本的内容分析[J]. 网络安全技术与应用, 2022(001):000.

[6] 全国信息安全标准化技术委员会秘书处. 网络安全标准实践指南—移动互联网应用程序（App）收集使用个人信息自评估指南[EB/OL]. (2020-07-22)[2023-05-08]. https://www.tc260.org.cn/front/postDetail.html?id=20200722134829.

[7] 工业和信息化部信息通信管理局. 关于侵害用户权益行为的APP（第一批）通报[EB/OL]. (2019-12-19)[2023-05-08]. http://www.cac.gov.cn/2019-12/19/c_1578298327550994.htm.

[8] 工业和信息化部信息通信管理局. 关于侵害用户权益行为的APP（SDK）通报（2023年第3批，总第29批）[EB/OL]. (2023-05-06)[2023-05-08]. https://mp.weixin.qq.com/s/vd_7dgo_yLAk46F74tiFpg.

[9] Staff F T C. Protecting consumer privacy in an era of rapid change a proposed framework for businesses and policymakers[J]. Journal of Privacy and Confidentiality, 2011, 3(1).

[10] 袁康. 规范化程度低、侵权风险高 App隐私协议现状调查[EB/OL]. (2021-08-19)[2023-05-08]. https://news.cctv.com/2021/08/19/ARTIZWiWyFSopK22fclsZRmd210819.shtml.

[11] Liu S, Zhao B, Guo R, et al. Have you been properly notified? automatic compliance analysis of privacy policy text with GDPR article 13[C]//Proceedings of the Web Conference 2021. 2021: 2154-2164.

[12] Yu Y, Cao J, Stojmenovic M, et al. Time-capturing dynamic graph embedding for temporal linkage evolution[J]. IEEE Transactions on Knowledge and Data Engineering, 2021.

[13] Andow B, Mahmud S Y, Whitaker J, et al. Actions speak louder than words: Entity-sensitive privacy policy and data flow analysis with PoliCheck[C]//29th USENIX Security Symposium (USENIX Security 20). USENIX Association, 2020: 985-1002.

[14] Nick G. Android Market Share and Other Statistics for 2023 [EB/OL]. (2023-02-11)[2023-05-08]. https://techjury.net/blog/android-market-share/

[15] Shrivastava G, Kumar P. Android application behavioural analysis for data leakage[J]. Expert Systems, 2021, 38(1): e12468.

[16] 刘晓建,彭玉坤.App合规性检测综述[J].计算机工程与应用,2023,59(03):1-12.

[17] 朱璋颖,陆亦恬,唐祝寿,张燕. 基于隐私政策条款和机器学习的应用分类[J]. 通信技术, 2020, 53(11):9.

[18] 姜盼盼.图书馆隐私政策合规性的依据与标准[J].图书馆建设,2019,No.298(04):79-86.

[19] 全国人民代表大会常务委员会. 中华人民共和国个人信息保护法[EB/OL].(2021-08-20)[2023-05-08].
http://www.npc.gov.cn/npc/c30834/202108/a8c4e3672c74491a80b53a172bb753fe.shtml.
[20] 张艳丰,邱怡. 我国移动阅读应用个人信息保护政策合规性测度研究[J]. 图书情报工作,2021,65(22):35-43.
[21] 赵波,刘贤刚,刘行,胡影. Android应用程序个人信息安全量化评估模型研究[J]. 通信技术,2020,53(08):2019-2026.
[22] Amaral O, Abualhaija S, Torre D, et al. AI-Enabled Automation for Completeness Checking of Privacy Policies[J]. IEEE Transactions on Software Engineering, 2021, 48(11): 4647-4674.
[23] Torre D, Abualhaija S, Sabetzadeh M, et al. An ai-assisted approach for checking the completeness of privacy policies against gdpr[C]//2020 IEEE 28th International Requirements Engineering Conference (RE). IEEE, 2020: 136-146.
[24] Mousavi Nejad N, Jabat P, Nedelchev R, et al. Establishing a strong baseline for privacy policy classification[C]//IFIP International Conference on ICT Systems Security and Privacy Protection. Springer, Cham, 2020: 370-383.
[25] Bhatia J, Breaux T D, Schaub F. Mining privacy goals from privacy policies using hybridized task recomposition[J]. ACM Transactions on Software Engineering and Methodology (TOSEM), 2016, 25(3): 1-24.
[26] Fan M, Yu L, Chen S, et al. An empirical evaluation of GDPR compliance violations in Android mHealth apps[C]//2020 IEEE 31st international symposium on software reliability engineering (ISSRE). IEEE, 2020: 253-264.
[27] 陈世敏. 中文可读性公式试拟[J]. 新闻学研究, 1971 (8): 181-225.
[28] 苗慧. 中外移动APP的个人信息保护研究[D].北京邮电大学,2021.
[29] 秦克飞.手机APP隐私政策的可读性研究[J].情报探索,2019(01):18-23.
[30] 朱侯,张明鑫,路永和.社交媒体用户隐私政策阅读意愿实证研究[J].情报学报,2018,37(04):362-371.
[31] Wilson S, Schaub F, Dara A A, et al. The creation and analysis of a website privacy policy corpus[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2016: 1330-1340.
[32] Poplavska E, Norton T B, Wilson S, et al. From prescription to description: Mapping the GDPR to a privacy policy corpus annotation scheme[C]//Legal Knowledge and Information Systems-JURIX 2020: 33rd Annual Conference. 2020.
[33] Zimmeck S, Story P, Smullen D, et al. Maps: Scaling privacy compliance analysis to a million apps[J]. Proceedings on Privacy Enhancing Technologies, 2019, 2019(3): 66-86.
[34] Sathyendra K M, Wilson S, Schaub F, et al. Identifying the provision of choices in privacy policy text[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017: 2774-2779.
[35] Zimmeck S, Story P, Smullen D, et al. Maps: Scaling privacy compliance analysis to a million apps[J]. Proc. Priv. Enhancing Tech., 2019, 2019: 66.

[36] Poplavska E, Norton T B, Wilson S, et al. From prescription to description: Mapping the GDPR to a privacy policy corpus annotation scheme[C]//Legal Knowledge and Information Systems-JURIX 2020: 33rd Annual Conference. 2020.
[37] Gupta S, Poplavska E, O'Toole N, et al. Creation and Analysis of an International Corpus of Privacy Laws[J]. arXiv preprint arXiv:2206.14169, 2022.
[38] 赵杨,严周周,沈棋琦.基于机器学习的医疗健康APP隐私政策合规性研究[J]. 数据分析与知识发现, 2022, 6(5): 112-126.
[39] 朱侯,吴子帅,韦秉东.基于BERT文本分类模型的APP隐私政策完整性评价研究[J].现代情报,2023,43(03):123-134.
[40] Zhao K, Yu L, Zhou S, et al. A Fine-grained Chinese Software Privacy Policy Dataset for Sequence Labeling and Regulation Compliant Identification[J]. arXiv preprint arXiv:2212.04357, 2022.
[41] 加利福尼亚州司法部. California Consumer Privacy Act (CCPA) [EB/OL]. (2023-03-10)[2022-05-09]. https://oag.ca.gov/privacy/ccpa.
[42] 弗吉尼亚州.Consumer Data Protection Act. [EB/OL].(2023-01-01)[2023-05-08]. https://law.lis.virginia.gov/vacodefull/title59.1/chapter53/
[43] CDC.Health Insurance Portability and Accountability Act of 1996 (HIPAA) [EB/OL].(2022-06-27)[2023-05-08].https://www.cdc.gov/phlp/publications/topic/hipaa.html.
[44] FTC. Graham-Leach-Bliley Act. [EB/OL].(2022-03-31)[2023-05-08]. https://www.ftc.gov/business-guidance/privacy-security/gramm-leach-bliley-act.
[45] Harkous H, Fawaz K, Lebret R, et al. Polisis: Automated analysis and presentation of privacy policies using deep learning[C]//27th {USENIX} security symposium ({USENIX} security 18). 2018: 531-548.
[46] Joachims T. Making large-scale SVM learning practical[R]. Technical report, 1998.
[47] Pregibon D. Logistic regression diagnostics[J]. The annals of statistics, 1981, 9(4): 705-724.
[48] Feller W. An introduction to probability theory and its applications, Volume 2[M]. John Wiley & Sons, 1991.
[49] Rabiner L. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition[J]. Proceedings of the IEEE, 1989, 77(2): 257-286.
[50] Hochreiter S, Schmidhuber J. Long Short-Term Memory[J]. Neural Computation, 1997, 9(8): 1735-1780.
[51] Schuster M, Paliwal K K. Bidirectional Recurrent Neural Networks[J]. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 1997, 1: 669-672.
[52] Kim Y. Convolutional Neural Networks for Sentence Classification[C]. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 2014: 1746-1751.
[53] Mikolov T, Chen K, Corrado G, et al. Efficient Estimation of Word Representations in Vector Space[C]. Proceedings of the International Conference on Learning Representations, 2013.
[54] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.
[55] Cohen J. A coefficient of agreement for nominal scales[J]. Educational and psychological measurement, 1960, 20(1): 37-46.

[56] Qian Y . CPP-130[EB/OL].(2023-05-06)[2023-05-08]. https://github.com/ylqianbj/CPP-130-Chinese-privacy-policy-dataset.
[57] Chen Y. Convolutional neural network for sentence classification[D]. University of Waterloo, 2015.
[58] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.
[59] Song Y, Shi S, Li J, et al. Directional skip-gram: Explicitly distinguishing left and right context for word embeddings[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). 2018: 175-180.

附　录

中文隐私政策标注手册

依托本研究在此前建立的隐私政策合规性评价指标体系，制定本手册。本手册用于建设“CPP-130”中文隐私政策数据集，标注过程需严格按照手册中的规定开展。

一、标注任务基本情况

本任务是标注隐私政策文件，隐私政策文件已经做好了分句，存储于一系列的csv文件当中。

“标注工具.xlsx”文件是一个标注工具，其第二列和第三列做好了标签的多级下拉菜单，可以利用这个简单的工具进行标注。

标注时，可以将csv文件中的隐私政策分句结果复制到标注工具文件中，利用多级下拉菜单，参照标注手册进行标注。标注完成后，将文件另存为一个csv文件，将文件名命名为与刚刚标注的csv文件相同即可。

标注时可能会遇到一个句子有可能可以标注为两个标签的情况，对于这种情况解决方案为：如果认为整个句子基本可以归入某一个标签，仅仅有少量的其他信息要归入另一个标签，即可忽略，标为最符合的标签即可；如果句子较为明确的属于两个标签，且可以较好的分割，则对句子进行分割，将分割后的句子粘贴到最后，分开进行标注；如果不能分割，则将整个句子粘贴在最后，两处分别标注为不同的标签。

二、标签体系

标签体系依据相关法律法规设计，共分为两级标签，下表中加粗显示的是一级标签，每个一级标签对应的列是其相应的二级标签。其中其他通用信息列没有二级标签。

第一方收集_使用	**第三方搜集_使用**	**与第三方共享_转让_公开**	**用户访问_编辑_删除的方法**	**数据安全**	**条款更改**	**面向特定人群用户条款**	**其他通用信息**
何时采集	第三方何时采集	何时与第三方共享	用户可以进行的操作	个人信息类型	更改原因	群体类型	
采集的目的	第三方相关实体	第三方相关实体	操作的数据类型	是否有交互	告知方式	用户选择	
采集的渠道	采集的目的	是否有交互	操作的途径	数据存储/跨境传输的目的	用户选择	供应商动作	
是否有交互	是否有交互	与第三方共享的目的	供应商动作	安全措施			
采集的个人信息类型	采集的个人信息类型	与第三方共享的方式	操作对用户造成的影响	数据存储期限			
用户的选择	用户的选择	与第三方共享的个人信息类型		发生安全事故的处理方式			
用户选择所造成的影响	用户选择所造成的影响	用户的选择		跨境传输			
对数据的使用	对数据的使用	用户选择所造成的影响					
使用的目的	使用的目的	第三方对数据的使用					
	对于第三方的约束	第三方使用数据的目的					
		对于第三方的约束					

三、标签内容

下面对于标签进行具体阐释。

1. 第一方收集_使用

第一方收集_使用是指，经用户同意隐私政策后，软件运营者所进行的收集和使用用户信息的一系列行为。

该属性包含以下9个二级属性：

(1) 何时采集：指软件运营者在用户进行何种操作时对用户进行的信息收集，例如注册、登录、使用特定功能等。

(2) 采集的目的：指软件运营者在用户进行何种操作时，因何种目的收集用户信息。例如，改善用户体验、向用户提供个性化服务、推广广告等。

(3) 采集的渠道：指软件运营者收集用户信息的渠道，例如用户自行填写、设备信息、第三方数据接口等。

(4) 是否有交互：指软件运营者在收集用户信息时是否需要与用户进行交互，例如用户勾选同意、输入验证码等。

(5) 采集的个人信息类型：指软件运营者在用户进行何种操作时，收集用户的具体何种个人信息，例如姓名、电话号码、地址等。

(6) 用户的选择：指软件运营者在收集用户信息时，是否给用户提供选择权，例如同意/拒绝选项、是否开启某个功能等。

(7) 用户选择所造成的影响：指用户在选择同意/拒绝后可能产生的影响，例如无法使用某个功能、影响个性化服务等。

(8) 对数据的使用：指软件运营者在收集用户信息后对数据的处理方式，例如加密、匿名化、存储等。

(9) 使用的目的：指软件运营者对收集到的用户信息的具体使用目的，例如个性化推荐、信息安全等。

第一方收集_使用的句子示例：你使用推荐通信录好友功能时，在获得你的明示同意后，我们会将你通信录中的信息进行高强度加密算法处理后，用于向你推荐通信录中的好友。

注意事项：第一方就是指软件运营者，通常会使用“我们”这样的表述。相比第三方收集/使用会明确声明“第三方”，第一方收集/使用可能不会明确指出，需要结合上下文进行判断。例如，如果上文提到了软件运营者收集某些信息，下文可能仍然在进一步阐述软件运营者相关的操作。

2. 第三方搜集_使用

第三方搜集_使用是指除了第一方之外的其他实体（如广告公司、分析公司等）通过用户使用本产品而采集到的个人信息的处理方式。

其具有10个二级属性：

(1) 第三方何时采集：第三方在用户进行何种操作时对用户进行的信息收集。

(2) 第三方相关实体：哪些第三方机构或实体将收集用户的个人信息。

(3) 采集的目的：第三方机构或实体为何目的收集用户信息。

(4) 是否有交互：用户是否与第三方机构或实体进行了互动。

(5) 采集的个人信息类型：第三方机构或实体在用户进行何种操作时，收集用户的具体何种个人信息。

(6) 用户的选择：用户是否可以选择拒绝第三方机构或实体对其个人信息的收集。

(7) 用户选择所造成的影响：如果用户拒绝第三方机构或实体对其个人信息的收集，将对用户使用本产品造成什么影响。

(8) 对数据的使用：第三方机构或实体如何处理、使用收集的个人信息。

(9) 使用的目的：第三方机构或实体收集个人信息的目的。

(10)对于第三方的约束：如何监督第三方机构或实体对个人信息的使用和保护。

第三方收集_使用的句子示例：为了更好地向您推荐广告和内容，我们可能会使用第三方的分析工具，它们将使用您的设备上的信息，如您的IP地址和搜索和浏览的网页和内容等，来收集有关您的信息。这些数据将帮助我们更好地理解您的兴趣和使用情况，从而提供更好的内容和广告。

注意事项：第三方收集/使用通常会明确声明“第三方”。用户应该仔细阅读隐私政策，了解哪些第三方机构或实体将收集他们的个人信息，并了解这些第三方机构或实体如何使用和保护他们的个人信息。

3. 与第三方共享_转让_公开

与第三方共享_转让_公开是指，第一方将用户信息与第三方分享、转让或公开的行为。

其具有11个二级属性：

(1) 何时与第三方共享：指第一方何时与第三方共享用户信息，例如在注册、交易、咨询等过程中。

(2) 第三方相关实体：指与第三方共享信息的公司、组织、个人等相关实体的名称和身份。

(3) 是否有交互：指共享信息的过程中是否涉及用户与第三方的交互。

(4) 与第三方共享的目的：指共享信息的目的和原因，例如推广、营销、交易等。

(5) 与第三方共享的方式：指共享信息的具体方式，例如传输、出售、租赁等。

(6) 与第三方共享的个人信息类型：指与第三方共享的个人信息类型，例如姓名、地址、联系方式等。

(7) 用户的选择：指用户是否可以选择是否与第三方共享信息。

(8) 用户选择所造成的影响：指用户选择是否与第三方共享信息所产生的影响，例如对服务的使用、优惠的享受等。

(9) 第三方对数据的使用：指第三方对共享的个人信息的使用方式。

(10) 第三方使用数据的目的：指第三方使用共享的个人信息的目的和原因。

(11) 对于第三方的约束：指第一方对第三方使用共享信息的行为所采取的约束和措施，例如签署保密协议、限制使用范围等。

与第三方共享_转让_公开的句子示例：我们可能会将您的信息共享给我们的合作伙伴，包括广告合作伙伴。我们只会共享必要的信息，并且这些合作伙伴将遵守本隐私政策和相关的数据保护法律。

注意事项：第三方共享_转让_公开通常会明确声明“第三方”的存在。

4. 用户访问_编辑_删除的方法

用户访问_编辑_删除的方法指的是用户对其个人信息进行访问、编辑或删除的方式和途径。

其具有5个二级属性：

(1) 用户可以进行的操作：指用户可以在哪些方面对其个人信息进行操作，如访问、编辑或删除等。

(2) 操作的数据类型：指用户可以对哪些类型的个人信息进行访问、编辑或删除操作。

(3) 操作的途径：指用户可以通过哪些方式进行访问、编辑或删除操作，如网站、APP等。

(4) 供应商动作：指供应商对用户个人信息访问、编辑或删除操作的响应和动作。

(5) 操作对用户造成的影响：指用户进行访问、编辑或删除操作所带来的影响，如可能影响用户使用服务的功能等。

用户访问_编辑_删除的方法的句子示例：用户可以登录账号，在个人中心进行个人信息的访问、编辑和删除操作。用户在进行访问、编辑或删除操作时，可能会影响到其使用部分功能。

5. 数据安全

数据安全指的是个人信息在存储、传输过程中的安全性。

其具有7个二级属性：

(1) 个人信息类型：指存储或传输的个人信息类型。

(2) 是否有交互：指数据在存储或传输过程中是否涉及用户与第三方的交互。

(3) 数据存储/跨境传输的目的：指数据存储或跨境传输的目的和原因。

(4) 安全措施：指为保障个人信息的安全所采取的技术和管理措施。

(5) 数据存储期限：指个人信息的存储期限。

(6) 发生安全事故的处理方式：指一旦发生个人信息安全事故后，公司将采取哪些应对和处理措施。

(7) 跨境传输：指个人信息是否会跨境传输。

数据安全的句子示例：我们通过使用安全的传输协议和数据加密技术来保护用户的个人信息。我们将在符合法律法规的要求下对用户的个人信息进行存储，并对个人信息进行严格的保密措施。在发生个人信息泄露等安全事件时，我们会立即采取相应的应对措施，并通知用户。

6. 条款更改

条款更改是指供应商在运营过程中可能会对服务条款进行修改、更新或调整的行为。这通常是由于法律法规、商业策略等因素引起的，因此需要告知用户并获得他们的同意。在条款更改时，供应商需要说明更改原因、告知方式以及用户选择是否接受新条款等相关信息。

其具有3个二级属性：

(1) 更改原因：在更新服务或法规的情况下，供应商可能需要修改条款。此外，供应商可能还需要更改条款以改进服务或解决问题。

(2) 告知方式：供应商应该在网站或应用程序上发布更改通知，并通过电子邮件或其他可靠方式将通知发送给用户。

(3) 用户选择：在发布更改通知后，用户有权选择是否继续使用服务。如果用户不同意更改，他们可以选择终止服务并注销帐户。如果用户继续使用服务，则视为已接受新条款。

7. 面向特定人群用户条款

面向特定人群用户条款是指供应商根据不同的用户群体需求和特点，对服务条款进行定制的行为。不同的用户群体可能有不同的需求和特殊条款，供应商需要根据用户的需求，为不同用户群体提供个性化的服务。这样可以更好地保护用户权益，并提高用户体验。在面向特定人群用户条款中，供应商需要说明群体类型、用户选择以及供应商动作等相关信息。

其具有3个二级属性：

(1) 群体类型：某些服务可能面向特定人群，例如儿童或残疾人。供应商应该提供相应的条款以保护这些用户的利益。

(2) 用户选择：这些用户也应该有权选择是否接受条款。供应商应该采取合适的措施来确保这些用户能够理解条款内容。

(3) 供应商动作：供应商可能需要采取额外的措施来确保这些用户的权益得到保护，例如限制某些功能或设置特定的隐私设置。

8. 其他通用信息

除了上述7类较为明确的数据实践之外，其他句子都可以分类为其他通用信息类。