学号： 205458

**东南大学专业学位型研究生**

**学位论文开题报告及论文实施计划**

院（系、所）  软件学院

学位类别 工程硕士

专业领域 软件工程

研究生姓名 李乐翔

指导教师（校内） 李幼平教授

指导教师（校外） 徐涛

开题报告日期 2022年3月

东南大学研究生院制表

填 表 须 知

1. 论文开题报告由研究生本人向审议小组报告并听取意见后，由研究生本人填写此表。
2. 论文开题报告填写完成后，必须经导师审批，通过后方能提交。

3、博士生应在第四学期内、硕士生应在第三学期内完成此开题报告。开题报告经研究生秘书在网上审核确认（硕士生至少半年、博士生至少一年）后方可申请答辩。

4、研究生开题前应填写查新报告。查新报告对专业学位博士作为必要环节。博士生查新工作可委托图书馆负责，也可在完成网络文献检索类研究生课程的学习或参加学校组织的网络文献检索培训后，自行组织查新检索，自行组织查新需要详细文献查新述评作为附件。自行查新报告须经导师审查后由开题报告审核专家组审核签字（或盖章）。硕士生开题查新参考上述办法，不作硬性要求。

5、本表一式两份，一份研究生自留放入本人“研究生档案材料袋”；一份由院（系、所）保存并归入院（系、所）研究生教学档案。

6、学位类别为：工程硕士；公共管理硕士；法律硕士（非法学）；工商管理硕士；建筑学硕士；风景园林硕士；临床医学硕士；公共卫生硕士；旅游管理硕士；会计硕士；国际商务硕士；资产评估硕士；工程管理硕士；艺术硕士；工程博士；医学博士等。

7、本表下载区：http://seugs.seu.edu.cn/3676/list.htm 。本表电子文档打印时用A4纸张，格式不变，内容较多可以加页。

一、学位论文开题报告

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 论 文 题 目 | 基于注意力机制的社交平台敏感文本及变体字检测 | | | | | | | | |
| 研 究  方 向 | 软件工程 | | | | | | | | |
| 题 目 来 源 | 国家 | 部委 | 省 | 市 | 厂、矿 | 自选 | 有无合同 | 经费数 | 备注 |
|  |  |  |  |  | √ |  |  |  |
| 题 目  类 型 | 理论  研究 | 应用  研究  **√** | 工程  技术 | 跨学科  研 究 | 其他 |  | | | |
| 开题报告内容（具体要求见《东南大学研究生论文选题和开题报告的原则和要求》） 一、选题依据与工程应用价值 1.1选题依据  受疫情影响，我国网民规模、互联网使用率和使用时长增长显著。2020上半年，电商直播、短视频和网络购物等应用的用户规模增长尤为显著，增长率分别高达16.7%、5.8%和5.5%。即时通信、搜索引擎、网络游戏和网络视频也保持增长，增长率在1%-5%之间[1]。根据中国互联网络信息中心（CNNIC）发布第48次《中国互联网络发展状况统计报告》[2]（以下简称《报告》）显示，截至2021年6月，我国网民规模达10.11亿，较2020年12月增长2175万，互联网普及率达71.6 %；随着互联网技术的不断发展，微博、微信朋友圈、QQ空间、抖音等主流社交平台的兴起，为人们提供了随时随地分享生活、获取资讯的途径，又一次改变了现代人的社交方式，第46次《报告》指出 ，截止2020年6月，微信朋友圈使用率为85.0%，QQ空间、微博使用率分别为41.6%和40.4%，如图1.1所示。微信朋友圈、微博等主流社交平台长期占据大部分流量，并通过不断丰富的即时通信、短视频、电商、本地生活等服务，构建出完善的流量闭环和服务生态，社交平台应用俨然成为互联网产业中的基础应用。  图1.1 2020.03和2020.06典型社交应用使用频率对比  人们在社交平台上分享生活的同时，一些不法分子利用互联网的匿名性和虚拟性，在社交平台中恶意传播散布违法违规的敏感文本信息来引流、博人眼球、甚至谋取个人利益，如：“薅羊毛、外挂辅助、加微刷单”等包含敏感词汇的文章标题和评论。这些借助互联网技术和网络平台，进行有组织、有目的、有分工且规模化的网络违法犯罪统称为网络黑灰色产业。其中，黑产用户是指存在违规违法行为，以及对互联网业务环境有直接危害的用户，如涉政言论、涉恐涉暴、低俗色情、电信诈骗等；灰产用户是指对互联网业务环境有间接危害，但不存在违法违规行为的用户，如：兼职刷单、恶意营销、广告引流、软件外挂等。对于黑灰产而言，能触达海量用户、传播曝光黑灰产信息最快、成本最低的手段便是文本，而社交平台拥有海量用户群体无疑是黑灰产业滋生的温床。近年来黑灰色产业引起互联网平台的广泛关注，2021年9月17日，微信发布《微信外部链接内容管理规范》，对诱导分享关注、利益诱惑、夸张语言胁迫、诱导拼团等第三方链接进行规范，为用户提供绿色、健康的网络生态环境[3]。黑灰产用户几乎根植于所有的互联网平台，呈现出黑灰色产业链的发展趋势，俨然成为了一套完整的生态。  要治理社交平台的黑灰色文本信息一般采用算法检测过滤或人工审核两种方式。黑灰色文本信息一般携带敏感文本，审核和检测黑灰色文本信息，实际是检测并识别出敏感文本。然而对于一个成熟的社交平台而言，每日的黑灰色内容文本超过百万，并混杂在每日十亿级别的文本数据中，通过人力全面审核是不可能完成的，而使用算法检测也要面临各种挑战。一、不同违规类型的敏感文本之间具有差异性：黑灰色文本信息的违规类型繁多，大致可以分为：涉政涉暴、低俗色情、涉赌涉毒、广告营销、虚假信息、标题党，违规类型不同，对应的敏感词库也有所不同，传统敏感文本检测算法是根据敏感词库匹配结果来判断文本中是否包含敏感信息；二、缺少明显的情感极性：与攻击性语言、仇恨言论、种族言论相比，黑灰色文本信息的感情色彩较弱，不具有较强的情感极性；三、中文敏感词具有多样性：与英文不同，中文是由拼音、偏旁部首构成，从而衍生出各种变体字：如同音异形、谐音字、拼音缩写、形近字、繁体字、偏旁拆分等诸多变化形式，如图1.2所示，黑灰产用户借助中文敏感词变体字与检测过滤算法进行对抗，恶意绕过算法检测，这为敏感文本检测带来了诸多挑战。黑灰色文本信息的传播严重影响绿色健康的网络环境甚至影响社会和国家的长治久安，因此如何在社交平台中准确、全面且快速地检测敏感文本是学术界和工业界亟待解决的关键问题。  图1.2 中文敏感文本变体字分类  敏感文本检测本质上属于分类任务，本课题计划使用一种基于Attention机制的深度学习模型进行敏感文本的分类任务；基于字音、字形等中文文本特征，对包含敏感词变体字的黑灰色文本信息处理与转换，以提高敏感文本分类的准确率，并完成敏感文本检测系统的开发与测试。  1.2工程应用价值  基于敏感词库匹配的传统中文违规文本的检测算法存在局限性：敏感词库匹配模型泛化能力较低，与中文敏感词变体字对抗性不足，难以有效应对敏感词文本内容不断变化的场景。通过研究中文敏感文本的拼音、汉字结构等特征，对敏感文本变体字进行预处理，并引入Attention机制的深度学习模型，更加关注敏感词本身，提高敏感文本检测的准确率，对于净化网络环境，打造清朗、健康、安全的网络空间，保护网民身心健康，以及维护社会稳定都有着重大的意义。 二、国内外研究现状与发展动态 敏感文本变体字转换和违规文本分类模型是检测中文敏感文本及其变体字的两个重要方向，但由于数据集和语言环境的限制，学术界和工业界只在敏感文本检测领域有一些初步尝试。随着神经网络、Attention机制以及各种预训练模型的兴起，逐渐成为自然语言处理领域的主流模型。本课题结合国内学者对中文敏感文本及其变体字转换方向的研究成果，对国内外学者在仇恨言论、虚假新闻、攻击性语言、种族言论等违规文本分类模型进行梳理。本章节将对研究现状和相关技术成果进行介绍。  2.1中文敏感文本及其变体字转换  国外的文本变体字主要是针对特殊符号与字母之间的相似性。Yoon等人[4]针对含有特殊符号的粗俗词，设计出基于近似字符串匹配的粗俗词过滤器，以及基于用户的粗俗词等级分级过滤。该方案给出了特殊字符与相似的英文字母之间的映射关系和匹配分数，如：完全匹配为1.0，“@”与“a”的匹配值为0.8；根据用户短期、长期内使用粗俗词的倾向总和，被使用的粗俗词数量与所有词汇数量之比为用户划分等级。对等级较高、粗俗词使用频繁的用户降低过滤的阈值，对等级较低、粗俗词使用不频繁的用户提高过滤阈值，解决斯肯普索问题（Scunthorpe problem），即因单词中的子字符串与粗俗词一致而被算法误伤（如：“classic”因包含“ass”而被误伤为“cl\*\*\*ic”）。  国内学者主要从拼音、字形角度研究汉字的变体字。Fu等人[5]提出一种基于变体字识别的敏感词检测方法。该方案采用编辑距离计算原始词汇与其汉语缩写（如“东南大学”简称“东大”）之间的相似度；根据同音异形词改变字符的首辅音或末音节的规律，提出一种基于拼音比较的方法对同音异形的词语进行检测，并在Internet Corpus(2008)搜狗网络语料中抽取2000作为数据集，检测效果优于基于敏感词库的检测方法、基于编辑距离的敏感词检测方法以及基于Bigram敏感词检测方法。周昊等人[6]提出一种综合考虑读音及字形特点的音形码汉字相似度改进算法，并通过添加精度参数的字典树完成敏感词检测。该方案根据汉语拼音的声母、韵母、声调，以及字形结构进行编码来计算相似度，并得到音码部分和形码部分在最终相似度计算中所占的贡献比；采用字典树的方法构建整个模式串库，并采用模糊匹配（设置匹配精度）的方式对目标串进行检索，在常用音近形近汉字组数据集中进行实验，与编辑距离和未改进的音形码方法相比具有更高的精度。  付聪等人[7]基于音码编码的基础上，对汉语拼音中易混拼音进行分组，设置相似性因子来表示拼音被替换需要付出的代价，判断敏感词与疑似敏感词之间的相似性，但文中相似性阈值的设定缺少理论支撑；根据词语的缩写过程中字序不变的特点判断敏感词与疑似敏感词之间是否存在缩写关系；人工对敏感词进行拆分，并利用汉字区位码对拆分后的敏感词进行编码，基于KMP算法判断敏感词与疑似敏感词之间是否匹配成功。另外，赵俊杰[8]提出一种基于关联规则的中文变体词识别算法，尽可能地分出多字词和短语，便于中文变体词识别；杜刚等人[9]针对同音字和形近字，提出一种拼音推理的方法，并通过Transformer进行转换，但未考虑到拆字的情况；谢鑫[10]利用五笔打字编码来解决形近字绕过机制，利用拼音替换方式来解决音近字绕过机制，将形体特征、音体特征、语义序列特征结合在一起，再引入注意力机制重点关注不良文字信息，最后建立循环神经网络检测是否为不良文本。  2.2违规文本分类模型  国外学者基于迁移学习的思想，设计了跨语言、跨领域的文本分类模型，在资源较少的语言中进行预测，为解决缺乏标准公开中文数据集的问题提供了指导性思路。Arkadipta等人[11]针对低资源语言环境的虚假新闻，提出了一种BERT模型；该模型隐藏层是基于12层Transformer架构，具有12个多头自注意力头，使用softmax进行文本分类。在5种语言环境下的虚假新闻数据集中实验，与FastText和ELMo相比，该方案在跨语言、跨领域、零语言训练样本的条件下具有良好泛化能力和较高的准确率。Ranasinghe等人[12]第一次将跨语言语境词嵌入应用于攻击性语言识别研究中。该方案利用现有的英文数据集SemEval-2019的任务6，提出了基于跨语言语境词嵌入和迁移学习的XLM-R[13]模型，在低资源语言的分类任务上表现显著。  敏感文本与仇恨言论、种族歧视、攻击性言论有一定的相似性，均属于违规文本范畴，并且关于仇恨言论等领域已经成为国内外学术界重点研究的课题，这对本课题具有一定的借鉴意义。Kavatagi等人[14]提出一种twitter数据中仇恨言论检测的语言模型，采用上下文感知嵌入模型Distil-BERT以及支持向量机用于仇恨言论分类，提高嵌入精度。Madukwe等人[15]针对仇恨言论检测任务，使用遗传算法搜索，选择和设计一个近似全局最优解的微调BERT架构。该方案将BERT层数、微调架构、CNN模型dropout、BiLSTM模型dropout等7个参数，进行二进制编码，根据参数的离散取值确定每个参数对应的位数，编码总长度为28位。实验使用遗传算法搜索出模型参数的近似解，可推广到其他文本分类任务中。D'Sa等人[16]认为仇恨言论（Hate Speech）、攻击性语言（Offensive Speech）、侮辱性语言（Abusive Speech）统称为有毒言论（Toxic Speech），它们之间没有明确的区别。该方案针对有毒言论，提出了使用基于特征的词嵌入FastText、BERT作为深度学习模型CNN、BiLSTM的输入，以及基于BERT微调模型。实验结果显示BERT微调模型的宏观平均F1值（Macro-average F1-measure）高于其他方法，可以用于任何其他类型的社交媒体评论。  虽然传统违规文本检测领域已经取得了一定的成果，但对于新领域的数据模型性能会变差。Li等人[17]认为COVID-19（新型冠状病毒肺炎）领域存在一些新型仇恨言论，如仇恨口罩、仇恨疫苗、仇恨老年人等，造成数据分布与传统仇恨言论不同，是导致传统模型变差的原因。因此提出了基于新领域数据再训练的COVID-HateBERT模型，以提高检测COVID-19相关仇恨言论的性能和模型的泛化能力。  随着国内社交平台的蓬勃发展，以及黑灰产业的泛滥，国内学者对中文敏感文本进行了研究。Yang等人[18]在YouTube平台识别中文敏感文本，提出一种将BERT与双向长短期记忆网络和条件随机场相结合的模型（BERT-BiLSTM-CRF）。具体来说，该模型使用BERT生成单词向量，克服了one-hot 与Word2Vec以及Glove等固定向量的嵌入式表示所带来维度稀疏性与一词多义等问题；双向LSTM模型避免了RNN梯度消失和梯度爆炸的问题，更注重上下文的语言环境；并添加CRF层输出标签。使用BIO格式对数据集进行标注，与CNN和BiLSTM-CRF模型相比，本方案有更高的F1值。Wang等人[19]从YouTube上选择了20个热门中文品，通过谷歌YouTubeV3 Api爬取视频下方的评论构建数据集。该方案使用BERT预训练单词和句子的嵌入，作为CNN的输入，构建中文敏感信息分类的BERT模型，与传统神经网络模型相比提高了词嵌入的泛化能力。由于当前对中文敏感文本检测的研究并不全面，并且与中文敏感文本及其变体字相关的数据集有限，本课题目标是对该领域进一步研究并填补空白。  2.3研究现状总结  综上所述，目前国内外学者在中文敏感词文本变体字转换和违规文本分类模型领域上已经取得了一定的成果，对本课题的研究具有一定的借鉴意义，但仍然有一些不足之处：  （1）利用中文变体字复杂性和多样性恶意绕过算法检测：现阶段的敏感词检测大多基于敏感词库，检测的对象是敏感词，属于词级分类word-level，其检测的准确性强依赖于敏感词库的容量；并且在实际场景中，敏感词并非以字、词的形式单独出现，而是利用变体字的复杂性和多样性隐藏在文本中，以恶意绕过算法的检测。  （2）汉字特征利用不充分：大多数针对中文文本的语言分类模型仅依靠分词、删除停顿词的方法构建词向量，忽略了汉字本身的特征，如汉语拼音、汉字字形结构、汉字笔画等，以上特征对理解中文语言有着深远的意义。  （3）针对中文敏感文本的分类模型检测能力不足：大多数违规文本（仇恨言论、攻击性言论等）分类模型是针对英文语言环境的，而汉字结构复杂多变，中英文语序也有较大差别；仇恨言论、攻击性言论一般具有情感极性，相较而言敏感文本与正常文本之间的界限模糊，这为检测中文违规文本带来了许多挑战。  针对以上不足，本课题收集并整理出适用于中文敏感词文本转化与识别的任务数据集，基于文本、图片特征的多模态中文变体字转换模型，并针对转换后的文本，提出一种改进掩蔽策略的敏感文本检测模型；最后设计并实现一种端到端的中文敏感文本检测系统。 三、研究目标与研究内容 3.1研究目标  针对汉字特征利用不充分的问题，本课题融合汉字字音、字形抽取变体字文本特征，通过图像形式抽取字形的空间特征；针对中文变体字恶意绕过算法检测的问题，本课题结合注意力机制建立中文变体字转换模型，提高模型与中文变体字之间的对抗性和泛化能力；针对中文敏感文本检测能力不足的问题，本课题提出一种改进掩蔽策略的中文敏感文本检测模型，进一步提高中文敏感文本的检测准确率。  3.2研究内容  结合研究背景和研究目标，本课题的研究内容可以概括为三点：融合多模态与注意力机制的中文变体字转换模型、改进掩蔽策略的中文敏感文本检测算法、端到端的社交平台敏感文本及变体字检测系统的开发与测试，如图3.1所示。  图3.1 研究目标和研究内容概览  下面将对这三点研究内容涉及的主要工作进行叙述：  **（1）融合多模态与注意力机制的中文变体字转换模型**  针对利用中文变体字恶意绕过算法检测和汉字特征利用不充分的问题。本课题提出一种融合汉字字音、字形、字符图像的词嵌入模型，并将注意力机制应用到中文变体字转换模型中，以期望在不影响文本正常内容的前提下，将变体字转换成正常文本。本课题基于Bi-LSTM抽取文本的序列特征，基于CNN抽取字符图像的字形特征和文本局部信息，并添加多头注意力机制，以便更好地对变体字进行转换，最后通过BLEU（N=4）指标来检测转换结果的准确性和流畅度。  **（2）改进掩蔽策略的中文敏感文本检测算法**  针对中文敏感文本分类模型检测能力不足的问题，本课题提出一种改进掩蔽策略的敏感文本检测方法，并将敏感文本检测归结为文本分类任务，以期能够在抽取中文敏感文本特征工作中能够取得更好的效果。该方法将经过（1）处理后得到转换后的文本作为检测模型的输入，使用BERT进行词嵌入，预训练过程中设计一种基于字音、字形的相似词替换的掩蔽策略得到特征矩阵，经过Softmax层输出后得到分类结果，使用F1评价指标，来判断中文敏感文本检测模型的准确性和有效性。  **（3）端到端的社交平台敏感文本及变体字检测系统**  基于以上理论研究和构思，本课题设计并实现一种端到端的社交平台敏感文本及变体字检测系统。该系统使用爬虫查询或社交网络官方提供的消息接口获取待检测的敏感文本，对于输入的敏感文本，抽取敏感文本字符、字音、字形特征，作为变体字转换模型的输入，转换后的文本送入敏感文本检测模型进行分类，对于已经被检测出的敏感文本，将敏感文本及其转换结果存储在数据库或数据仓库中，从而构建敏感文本数据库。 四、实施方案及可行性分析 4.1融合多模态与注意力机制的中文变体字转换模型  4.1.1中文敏感文本数据集  本课题主要有两个研究内容：中文敏感词及其变体字转换模型和中文敏感词检测模型，并且对数据集有一定的特殊要求，要求包含中文敏感词和中文敏感词变体字。学术界研究人员利用爬虫技术爬取新浪微博、YouTube等社交平台文本数据来构建数据集；工业界和社交平台厂商虽然能获取到数据，但此类数据与用户个人信息和商业机密强相关，一般不对外公开。正是由于该领域存在数据壁垒，缺少标准公开数据集，因此需要重新爬取与收集。  本课题从社交媒体平台上收集整理一份中文敏感词文本数据集，数据由腾讯科技（北京）有限公司提供，数据出自腾讯企鹅号平台文章标题。数据集总体数据量为16000条，正负向数据量比例大致为6:4，其中负向数据为敏感文本，部分负向数据包含中文变体字，正负向数据标签分别用0,1区分。数据基本信息如表4-1所示，数据示例如图4.1所示。  表4-1 企鹅号敏感文本数据集基本信息   |  |  |  | | --- | --- | --- | | 数据类型  基本信息 | 正向数据 | 负向数据 | | 数据量 | 9572 | 6428 | | 数据列信息 | content、label | content、label | | 数据标签 | 0 | 1 | | 数据长度范围 | [5,68] | [5,705] | | 数据平均长度 | 37.768 | 44.987 |   图4.1 企鹅号敏感文本数据集  目前存在一些研究机构和互联网厂商开展了中文敏感文本检测相关实验和赛事。2021年8月中国计算机学会、大数据协同安全技术国家工程实验室、360集团联合举办“面向黑灰产治理的恶意短信变体字还原”赛事[[1]](#footnote-1)。主要任务是对训练集中的短信样本进行分析，采用深度学习建模的方法将测试集中新出现的短信变体字还原为正常信息文本，即不含有变体字、干扰字符，所有变体字部分应使用常见简体汉字、字符来表示，同时需要保证不包含变体字的正常文本不受影响。大赛提供了一份公开中文恶意文本数据集，分为训练集和测试集。训练集给出原始恶意数据和人工还原的恶意文本，即人工标注文本，如表4-2所示，训练集和测试集数据示例分别如图4.2、4.3所示。  表4-2 面向黑灰产治理的恶意短信变体字还原”赛事恶意文本数据集   |  |  |  | | --- | --- | --- | | 数据类型  基本信息 | 训练集 | 测试集 | | 数据量 | 11968 | 49557 | | 数据列信息 | id、原始恶意文本、人工标注文本 | id、原始恶意文本 | | 数据标签 | 人工标注文本 |  | | 数据长度范围 | [4,524] | [1,230146] | | 数据平均长度 | 53.922 | 46.828 |   图4.2 恶意文本训练集  图4.3 恶意文本测试集  4.1.2数据预处理  **（1）数据清洗**  对原始数据进行去重、清洗，保留文本内容中的中文、英文、数字、部分特殊符号：1. 长度处理：删去长度小于等于5的数据；2. HTML标签处理：删除HTML标签，如<br>、<h1>、&nbsp；3. 符号处理：将逗号、逗号替换为空格；由于一些敏感词变体字中包含特殊符号（如“加微信”变体字为“+V”），需要保留数据中的括号、加号、空格、顿号、引号、小数点。  **（2）数据转换**  首先将敏感文本做字符级character-level分词，提取每个中文字符对应的字符id；（1）针对中文字音特征，采用pypinyin库获取每个中文字符对应的拼音和声调；（2）针对中文字形结构特征，采用hanzi\_chaizi库将中文字符按照偏旁部首进行拆分；（3）提取中文字形结构特征除了采用偏旁部首外，还有一种基于汉字四角号码的方法，四角号码是汉字词典最常用的检字方法之一，主要依据字形特征对汉字进行编码[6]，采用char\_featurizer库获取中文字符的四角号码特征；（4）中文字符其他特征，如笔画、五笔，其中通过pywubi库提取中文字符的五笔特征；（5）字形图像特征，将训练集语料中的汉字转换为不同字体风格的图片（如繁体、篆书）。以变体字文本“哇，槍 虹 笣，叁重忧蕙，葙請 BB6666 点C0M。DRN”为例，其特征嵌入形式如图4.4所示  图4.4 针对中文字音、字形特征提取  **（3）基于字音、笔画增强的文本对齐方法**  中文变体字转换任务类似机器翻译，属于字符级别character-level，要求输入的文本与输出的标签长度一致，即输入的每个字符，都需要有相应的分类标签与之对应，因此本课题设计一种基于字音、笔画的文本对齐方法，通过求出两个字的拼音、笔画的最长公共子串的长度，使用该长度除以较长字符串的长度得到音近、形近得分，根据设定的阈值来判断是否为音近、形近字，从而使同音字、形近字与转换后的正常文本一一对应，对齐方法过程如图4.5所示。  图4.5 基于字音、笔画的文本对齐方法  4.1.3基于Attention的中文变体字转换模型  黑灰产用户利用变体字、干扰字符恶意绕过平台算法检测的情况越来越多，由于变体字变换方式多，变换速度快，单纯通过规则进行变体词发现的效果有限，配套人工审核成本高且具有滞后性，因此需要一个转换敏感文本变体字的模型。本课题提出一种基于字音、字形、字符图像特征的词嵌入和注意力机制的中文变体字转换模型，转换模型框架如图4.6所示。  图4.6 中文变体字转换模型框架  **（1）基于CNN抽取字形的空间特征**  汉字本身为象形文字，随着历史的发展和书写的便利不断演化成如今的简体汉字，简体汉字也因此存在文字象形特征丢失的问题。例如“人”和“入”在字形、笔画、四角编码均相同，但在篆书中有明显差别[20]，如图4.7所示。因此本课题采用不同历史时期字体风格图片（繁体和篆书），与简体字相比这些历史字体会携带更多的象形信息，并基于CNN抽取字形的空间特征，模型结构如图4.8所示。  图4.7 “人”“入”字形特征  图4.8 卷积神经网络结构图  其中卷积层计算方法：输入特征图的尺寸为：，分别表示输入特征图的高、宽、通道数；通过公式（1）（2）（3），得到输出特征图的尺寸，其中输出通道数为，正方形卷积核的边长为，步幅(stride)为，补零的行数和列数(padding)为。  （1）  （2）  （3）  **（2）基于Bi-LSTM抽取文本序列特征**  由于变体字的复杂性和多样性，导致只关注当前字符无法判断出正常字符，因此需要将当前字符与前后的几个字一起组成词语或者固定搭配，故本课题利用Bi-LSTM抽取文本的序列特征，更好地表征文本上下文信息。与传统循环神经网络（RNN）相比，长短时记忆模块（Long Short-Term Memory）可以避免长期依赖问题，即LSTM模块，但标准的LSTM网络以时间顺序处理序列，忽略了未来的上下文。双向LSTM网络[21]通过引入第二个层来扩展单向的LSTM网络，在这个层中隐藏的隐藏连接以相反的时间顺序流动，因此该模型能够利用过去和未来的信息。  以LSTM模块为例，主要包含了三个门（forget gate、input gate、output gate）与一个记忆单元（cell），如图4.9所示，、和分别为遗忘门、输入门、输出门，用Sigmoid层表示。为记忆单元cell，LSTM可以通过门控单元对cell添加和删除信息。具体计算步骤如下    图4.9 LSTM模块结构图  遗忘门： (4)  输入门： (5)  候选记忆单元： (6)  当前时刻记忆单元： (7)  输出门： (8)  模型输出： (9)  **（3）注意力机制**  变体文本常常利用汉语中多音字来混淆视听，且变体文本中的每个字符对转换结果的贡献度不同。如“轧（zhà，yà）”，对于变体文本“边玩还交友抖地註,轧今花/牛\*牜”，当“轧今花”的“轧”字音嵌入结果为“yà”时，会干扰转换结果，因此需要模型关注文本中“抖地註”（斗地主），“牛\*牜”（牛牛，一种赌博牌类游戏），本课题设计一种注意力机制对文本序列特征和空间特征的关键部分进行捕获。  注意力机制即Attention mechanism[22]，在NLP上最早是被用于seq2seq的翻译类任务中，克服了Seq2Seq模型中固定维度大小的语义向量造成模型难以处理长文本的问题[23][24]。Attention机制可以更好的表征文本，使训练出的模型具有更好的表现。  图4.10 Scaled Attention和Multi-Head Attention模型结构  Attention机制包括Scaled Attention和Multi-Head Attention，如图4.10所示。Scaled Attention的输出矩阵可以按照式（3）进行计算，其中，Q、K、V均为64维向量，分别表示Query、Key、Value，是模型计算的中间变量，有助于计算注意力。  （10）  Multi-Head Attention则是对Attention进一步的改善，它拓展了模型关注不同位置的能力同时为attention层提供了多个表示子空间，可以按照式（4）进行计算，其中为式(5)的第个时间点计算的值，Concat表示拼接向量。  （11）  最后基于Softmax层选择出最高概率对应的词，即为这个时间步的输出的单词。  4.2改进掩蔽策略的中文敏感文本检测算法  除了变体字、干扰字符为敏感文本检测带来的挑战之外，黑灰产用户与普通用户之间、黑灰产文本内容与正常文本内容之间的界限越来越模糊。传统TextCNN无法学习到文本之间的关联性，RNN虽然可以学习到语句上下文之间的关系，但由于敏感文本与正常文本内容界限模糊难以区分，RNN效果不佳。因此本课题基于BERT预训练模型，并设计一种基于字音、字形的相似词替换的掩蔽策略应用到敏感文本检测当中，以缩小预训练和微调之间的差距。模型结构如图4.11所示。  图4.11 敏感文本检测模型结构  4.2.1基于BERT的词嵌入  BERT全称Bidirectional Encoder Representations from Transformers，是一种基于Transformer的双向编码器表示，2018年由Google提出[25]，已经成为最近几年最流行的NLP解决方案的范式。BERT的训练过程可以简述为：预训练（pre-training）和微调（finetune）。传统基于Word2Vec的词嵌入模型是对词的静态表征，不能解决一词多义的问题[26]，但汉语本身存在多音字，且对于中文敏感文本来讲，有许多一词多义和行业“黑话”的场景，如“品茶”、“韭菜”、“打渔”等，因此本文采用基于微调的BERT预训练模型进行词嵌入，得到特征矩阵。  经过中文变体字转换模型处理后得到文本结果，此类文本数据本身具有长度短、包含网络用语、句子结构不完整和混杂英文数字的特点，与BERT-Base[[2]](#footnote-2)训练中使用的中文语料库有较大差异，因此需要使用中文敏感文本数据对BERT-Base模型进行二次预训练，从而提高模型词嵌入效果。  C:\Users\SLZZ\AppData\Roaming\Tencent\Users\839908190\TIM\WinTemp\RichOle\_@URUY~%(LK2YB$_Z7)Q0PX.png0BERT的输入的编码向量是3个嵌入特征的单位和，分别为词嵌入Token Embedding、位置信息嵌入Position Embedding和句子信息嵌入Segment Embedding。词嵌入使用的是WordPieces作为最小的处理单元（采用的是wordpiece算法分词）：token，而不是使用单词本身；位置信息嵌入指将单词的位置信息编码成特征向量；句子信息嵌入指当前词所在句子的索引向量。BERT的输入特征如图4.12所示，其中[CLS]表示该特征用于分类模型，对非分类模型，该符合可以省去。[SEP]表示分句符号，用于断开输入语料中的两个句子。  图4.12 BERT的输入特征  基于Transformer的双向编码器表示是BERT模型的核心，也是解决一词多义问题的的关键。BERT-BASE包含12层encoder，每个encoder中都含有self-attention层，如图4.13所示。不同self-attention层中的权重矩阵也有所不同，因此对于一个词，在不同的上下文情况下它最后对应的12层Transformer的输出结果是不一样的，输出的词向量也不同。  图4.13 BERT-BASE12层Encoder结构  4.2.2一种相似词替换的掩蔽策略  BERT预训练模型WordPiece[27]的方法将单词的词根和词缀分开，如“playing”经过WordPiece处理后变为“play”和“ing”，有效减少了词表的数量，并通过掩蔽策略对单词片段进行掩蔽，即用[Mask]替换被掩蔽的单词，但由于掩蔽策略会引入[Mask]，而[Mask]从未在下游微调任务中出现，导致预训练和微调两个阶段存在差异，影响模型效果。与英文不同的是，中文预训练模型是对中文字符进行掩蔽，但仍然会有上述的问题，传统掩蔽策略如图4.14所示。  图4.14 传统掩蔽策略  业界也针对掩蔽策略提出了一些改进方法：百度公司提出ERNIE[28]，旨在学习通过知识掩蔽策略增强的语言表示，包括实体级掩蔽和短语级掩蔽，为模型提供更多的先验知识；Cui等人提出MacBERT[29]，基于Word2Vec[30]相似度计算的掩蔽策略，将掩蔽策略改为纠错模型；Liu等人提出RoBERTa[31]，采用动态掩蔽策略，每次向模型输入一个序列时都会生成新的掩码模式，增加模型的鲁棒性。  与Cui等人提出MacBERT[29]不同，本文设计一种基于字音、字形的相似词替换的掩蔽策略，以减少掩蔽标识[Mask]对下游微调任务的影响，也更加适合中文敏感文本检测场景。在上游数据预处理阶段中，构建一张基于字音、字形的词表，依据词表进行相似度计算，取相似度最高的字词对被掩蔽词进行替换。如图4.15所示。  图4.15 改进的掩蔽策略  4.3敏感文本检测系统的开发与测试  本课题计划设计并实现敏感文本检测系统，便于审核人员对敏感文本进行复查，以验证模型效果。系统共包括三个模块：用户发文模块、模型检测模块以及审核模块，敏感文本检测系统架构如图4.16所示。  图4.16 敏感文本检测系统架构  整个敏感文本检测系统流程如下：  （1）用户发文模块包含注册功能、登录功能、发文功能和评论转发功能。用户通过客户端访问社交平台页面，完成注册、登录后，用户即可在平台中发送推文，进行评论；  （2）模型检测模块包含变体字转换功能和敏感文本检测功能。针对用户发文内容，使用文本检测模块对用户推文内容进行转换和分类，根据分类结果将敏感文本存入敏感文本数据库并展示在审核平台上。  （3）审核模块包含审核功能和指标统计功能。审核人员浏览平台页面、查阅敏感文本数据库，以加大对敏感文本的整治力度；平台计算并展示出每日敏感文本审核量、通过量，否决量等指标。  4.4可行性分析  从技术、平台、时间三个角度，本文的可行性分析如下。  （1）技术可行性  自然语言处理技术日趋成熟，可供参考的分类模型、翻译模型的开源代码也越来越多，因此本课题的模型构建工作变得清晰易行。随着国家加大对网络环境的监管粒度，黑灰产业和敏感文本也受到了广泛关注，包括社交平台数据和CCF大数据与计算智能大赛提供的“恶意文本数据集”，打破了数据壁垒。  （2）平台可行性  本课题组隶属于教育部重点实验室，拥有充足的软、硬件资源，能够满足本课题涉及的研究和开发工作需求。此外本课题的相关工作还受到老师、同学的指导和帮助，因此本课题在平台方面是可行的。  （3）时间可行性  本课题的研究工作将在开题后开始，从开题至毕业有15个月的时间，具体工作分为文献整理、算法研究与实验、原型系统开发与测试以及论文撰写4个部分，各项工作将稳步推进，并按时汇报工作进度。 五、参考文献  * 1. 百度时代网络技术（北京）有限公司.2020网络黑灰产犯罪研究报告[EB/OL].北京.2020.https://wenku.baidu.com/view/61fe9cc2effdc8d376eeaeaad1f34693dbef1082.html   2. 中国互联网络信息中心. 第48次中国互联网络发展状况统计报告[EB/OL]. 北京: 2021.8 http://images.mofcom.gov.cn/lczx/202108/20210827170205314.pdf   3. 腾讯科技有限公司.微信团队.微信外部链接内容管理规范[EB/OL].2021.09. https://weixin.qq.com/cgibin/readtemplate?t=weixin\_external\_links\_content\_management\_specification   4. Yoon T, Park S Y, Cho H G. A smart filtering system for newly coined profanities by using approximate string alignment[C]//2010 10th IEEE International Conference on Computer and Information Technology. IEEE, 2010: 643-650.   5. Y Fu, Y Yu, X Wu. A Sensitive Word Detection Method Based on Variants Recognition[C] 2019 International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI). 2019: 47-52.   6. 周昊, 沈庆宏. 基于改进音形码的中文敏感词检测算法[J]. 南京大学学报：自然科学版, 2020, 56(2):8: 270-277.   7. 付聪, 余敦辉, 张灵莉. 面向中文敏感词变体字的识别方法研究[J]. 计算机应用研究, 2019, 36(4): 988-991.   8. 赵俊杰. 一种基于关联规则的中文变体词识别算法[J]. 重庆理工大学学报: 自然科学, 2018 (3): 178-185.   9. 杜刚,朱艳云,张晨,杜雪涛.变体垃圾短信翻译技术研究[J].电信工程技术与标准化,2020,33(07): 83-88.   10. 谢鑫. 基于深度学习的不良信息检测技术的研究[D]. 电子科技大学.   11. De A, Bandyopadhyay D, Gain B, et al. A Transformer-Based Approach to Multilingual Fake News Detection in Low-Resource Languages[J]. Transactions on Asian and Low-Resource Language Information Processing, 2021, 21(1): 1-20.   12. Ranasinghe T, Zampieri M. Multilingual offensive language identification for low-resource languages[J]. Transactions on Asian and Low-Resource Language Information Processing, 2021, 21(1): 1-13.   13. Conneau A , Khandelwal K , Goyal N , et al. Unsupervised Cross-lingual Representation Learning at Scale[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020:1-12.   14. Kavatagi S, Rachh R. A Context Aware Embedding for the Detection of Hate Speech in Social Media Networks[C].2021 International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON). IEEE, 2021: 1-4.   15. Madukwe K J , Gao X , Xue B . A GA-Based Approach to Fine-Tuning BERT for Hate Speech Detection[C].2020 IEEE Symposium Series on Computational Intelligence (SSCI). IEEE, 2020: 2821-2828.   16. D'Sa A G, Illina I, Fohr D. Bert and fasttext embeddings for automatic detection of toxic speech[C]//2020 International Multi-Conference on:“Organization of Knowledge and Advanced Technologies”(OCTA). IEEE, 2020: 1-5.   17. Li M, Liao S, Okpala E, et al. COVID-HateBERT: a Pre-trained Language Model for COVID-19 related Hate Speech Detection[C]//2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA). IEEE, 2021: 233-238.   18. Yang Y , Shen X , Wang Y . BERT-BiLSTM-CRF for Chinese Sensitive Vocabulary Recognition[C] International Symposium on Intelligence Computation and Applications. Springer, Singapore.2019: 257-268.   19. Wang Y , X Shen, Yang Y . The Classification of Chinese Sensitive Information Based on BERT-CNN[J].2019:269-280.   20. Meng Y, Wu W, Wang F, et al. Glyce: Glyph-vectors for chinese character representations[J]. Advances in Neural Information Processing Systems, 2019, 32:1-12.   21. Zhou P, Shi W, Tian J, et al. Attention-based bidirectional long short-term memory networks for relation classification[C]. Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers). 2016: 207-212.   22. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30:1-15.   23. Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks[J]. Advances in neural information processing systems, 2014, 27: 1-9.   24. Cho K , Merrienboer B V , Gulcehre C , et al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation[J]. Computer Science, 2014: 1724-1734.   25. Devlin J , Chang M W , Lee K , et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[J]. 2018:1-16.   26. Mikolov T , Chen K , Corrado G , et al. Efficient Estimation of Word Representations in Vector Space[J]. Computer Science, 2013: 1-12.   27. Wu Y , Schuster M , Chen Z , et al. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation[J]. 2016: 1-23.   28. Sun Y , Wang S , Li Y , et al. ERNIE: Enhanced Representation through Knowledge Integration[J]. 2019: 1-8.   29. Cui Y, Che W, Liu T, et al. Pre-training with whole word masking for chinese bert[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 3504-3514.   30. Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[J]. Advances in neural information processing systems, 2013, 26.   31. Liu Y , Ott M , Goyal N , et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach[J]. 2019: 1-13.   研究生签名  2022年 03月 12日 | | | | | | | | | |

二、学位论文工作实施计划

（一）论文的理论分析与硬件要求及其预期达到的水平与结果

|  |
| --- |
| **实验环境：**  硬件环境：PC机一台（Inter Core I5 CPU, 8G内存, 64位操作系统）  **预期成果：**  设计端到端的社交平台敏感文本及变体字检测系统，能够检测从网页上获得的信息是否为敏感文本，并有较高的检测精度。 |

（二）论文工作进度与安排

|  |  |  |
| --- | --- | --- |
| 起讫  日期 | 工 作 内 容 和 要 求 | 备 注 |
| 2022年3-7月 | 搜集文献资料，研究文本转换和分类技术 |  |
| 2022年8月 | 研究文本、图片特征嵌入原理及其实现 |  |
| 2022年9-10月 | 研究机器翻译和注意力机制相关算法及实现 |  |
| 2022年11-12月 | 研究预训练模型的设计、改进与实现 |  |
| 2023年1月 | 原型系统的开发与测试 |  |
| 2023年2-3月 | 整理撰写学位论文 |  |
| 2023年4月 | 论文修改和完善 |  |
| 2023年5月 | 论文答辩准备 |  |
| 学校指导教师对开题报告的综合意见 | 以《基于注意力机制的社交平台敏感文本及变体字检测》为题，以中文变体字转换模型和中文敏感文本检测模型展开研究，参考文献内容详实，对课题相关的前沿成果进行了充分的调研，选题符合专业培养目标，具有一定的研究价值，课题方案明确，实验设计合理，能够在规定时间完成，同意其开题。  指导教师（签字）  2022年 03月 12日 | |
| 校外指导  教师对开题报告的综合意见 | 《基于注意力机制的社交平台敏感文本及变体字检测》一题，紧扣专业方向，贴近现实生活，与实习体会结合，具有现实意义。该生在前期通过查阅文献，对研究内容所涉及的研究领域进行了较为全面的调研。该课题研究方向明确，实验方案可行，参考文献详实，文章结构完整，同意其开题。    指导教师（签字）  2022年 03月 12日 | |
| 开  题  报  告  审  议  情  况  记  录 | １、审议小组成员（硕士至少5人，博士5－7人，其中1人须为校外导师）：  组长：  成员：  ２、审议小组意见  ３、投票表决结果  审议小组出席 人；通过 人；不通过 人。  开题报告质量 （优、良、中、通过）  ４、审议小组组长（签名）  审议小组成员（签名）  年 月 日 | |
| 院（系、所）意见：  院（系、所）负责人签名（或印章）  年 月 日 | | |
| 备注： | | |

1. 面向黑灰产治理的恶意短信变体字还原大赛. https://www.datafountain.cn/competitions/508/ [↑](#footnote-ref-1)
2. Bert中文预训练模型. https://s3.amazonaws.com/models.huggingface.co/bert/bert-base-chinese.tar.gz [↑](#footnote-ref-2)