基于Transformer的语言暴力检测系统的实践研究

陈冠男
重庆理工大学　重庆市　401135

摘　要：随着互联网技术的不断发展，人们便捷交流下不断产生了因言语摩擦带来的语言暴力。但传统检测方法存在语义、上下文分析等方面不足。本文聚焦于xlm-roberta模型对于语言暴力检测的实现，主要基于Transformer的语言暴力检测系统着力于数据集的预处理、xlm-roberta模型的构建与训练以及基于tkinter的用户交互界面的搭建。系统能实现语言暴力检测效果，具备很高的鲁棒性，能为网络语言检测研究提供一定的方向。

关键词：NLP；Transformer；bert；xlm-roberta；text classification

绪论

当今数字化时代，社交媒体和网络平台已成为人们交流沟通、获取信息和表达观点的重要渠道。然而，随着用户的不断增加以及信息传播的日益便捷，网络环境中语言暴力（如仇恨性言论、冒犯性语言、威胁性话语等）现象愈发严重，引发了广泛的社会关注[1]。这些不良内容不仅对个人的心理和情感造成极大伤害，还严重破坏网络环境的和谐稳定，甚至可能引发线下的冲突与不良行为，对社会秩序产生负面影响[2]。尽管已有诸多研究致力于语言暴力检测，但目前的检测技术仍面临诸多挑战。不同语言的语法结构、词汇表达和文化背景等差异，使得统一的检测模型难以在各种语言上都取得理想的效果[3]，这些问题限制了语言暴力检测系统的有效性和可靠性，难以满足社会对净化网络环境的迫切需求[4]。本研究将聚焦于xlm-roberta模型对于语言暴力检测的实现。

1　系统设计

1.1 系统总体设计

本系统计划采用模块化架构，以用户交互模块为中心，主要及次级的语言暴力检测（分类）网络的构建为基础，搭建成整个语言暴力检测系统。其数据流向为：用户在交互模块输入待检测的文本，待检测文本由用户模块调用两个语言暴力检测模块进行检测，检测流程完成之后两个检测模块返回检测结果给用户交互模块，由交互界面处理检测结果后将检测结果显示在用户交互界面中的结果展示框中。

1.2 用户交互界面模块设计

本模块负责实现用户与整个语言暴力检测系统的整体的交互。具体流程是：接收用户输入的用于语言暴力检测的文本，并将用户输入的文本通过某种方式送入到语言暴力主分类网络和语言暴力次级分类网络中进行文本的检测，在得到两个分类网络的检测结果之后做出判断，并将检测结果通过展示框展示出来。用户交互界面模块与语言暴力主分类网络和语言暴力次级分类网络这两个网络的通信详细设计。

1.3 语言暴力主分类网络设计

语言暴力主分类网络用于：首先判断用户输入文本是否属于语言暴力，这是一个二分类问题，故该网络分类头设定为2个神经元用于输出。结合用户可能会遇到的语言暴力内容多为网络论坛、聊天在线平台、游戏平台等各种网络平台中的内容，而网络平台中的内容多为中英混杂甚至与是多语言混杂、多语言与数字及各种语言符号的混合内容，故采用能适用于多种语言、数字、符合混合的xlm-roberta网络来作为本系统的语言暴力检测器。其次，考虑到开展所拥有的算例比较紧张，故采用xlm-roberta-base网络，该网络相较于xlm-roberta-large网络参数量更少，计算时间更少，更有利于本次系统搭建时的调试、推进。

1.4 语言暴力次级网络设计

该网络与语言暴力主分类网络一样，考虑到用户可能会遇到的语言暴力内容以及本次我所拥有的算力资源，同样使用xlm-roberta-base网络；但该网络用于语言暴力次级分类，也就是判断用户输入的文本到底属于更细分的哪一类语言暴力（如种族歧视、地域歧视、性别歧视等细分分类），属于多分类问题，且对于这种多分类问题，数据集不容易收集，常会遇到小数据样本的情况，故考虑在xlm-roberta-base网络后添加自定义分类头，如自定义拥有两层隐藏层，并且带有dropout、层归一化、

残差链接的自定义分类头，具体的参数设定根据任务实际推进情况设定。

2 系统实现

主要分为以下几个部分：系统总体架构设计、数据预处理阶段、文本中语言暴力特征提取网络的构建、模型训练阶段、模型测试结果以及用户交互界面模块的实现。

2.1 系统总体架构实现概览

编写预处理模块群用于数据各项预处理操作、采用XLM-RoBERTa-base网络作为特征提取器、编写训练模块用于网络训练、编写测试模块用于模型训练效果测试、使用tkinter搭建用户交互界面用于用户便捷使用。

2.2 主网络构建

2.2.1 数据预处理

在数据预处理阶段存在多项预处理操作，各项操作如下：编写violence-generate.py和non-violence-generate.py文件用于生成一定规模的语言暴力和非语言暴力的数据集扩充。对于violence-generate.py也就是语言暴力生成这个模块，选择使用多维度词库配合不同类型的暴力模板库批量生成符合规则的语言暴力语句，多维度词库和暴力模板库，此模块的伪代码（略），保存为violence_data.csv。violence-generate.py生成的语言暴力语句样例（略）。对于non-violence-generate.py也就是非语言暴力生成这个模块，同样选择使用多维度词库配合不同类型的非暴力模板库批量生成符合规则的非语言暴力语句，多维度词库和非暴力模板库（略），此模块的伪代码（略），保存为no_violence_data.csv non-violence-generate.py生成的非语言暴力语句样例（略）。

（1）编写xieyintihuan.py文件伪代码（略），编写json2csv.py，toUTF-8.py，txt2csv.py，clean.py用于清洗收集到的不同的格式各异的数据集。

（2）编写swapping.py，label_num.py，merge.py，split.py文件用于数据集格式的调整、各个小数据集合并为一个整体数据集、按指定格式拆分为用于训练tran.csv，val.csv，test.csv三个文件。

2.2.2 特征提取网络构建

采用XLM-RoBERTa-base网络作为本系统主要的语言暴力特征提取器，选择XLM-RoBERTa是因为语言暴力检测的对象多为网络数字平台如数字社交平台、游戏平台、在线聊天平台、在线论坛等，对于这些场景，可能出现的语言暴力多为现代网络用语如热梗、谐音词、meme等，这类对象多出现中英文混杂，甚至更多语言或数字及各类符号的混杂，选用广泛使用的bert对于这类多语言混杂的场景效果不好，故采用XLM-RoBERTa这个适用于多语言的网络。并且，结合实际所拥有的紧张的算力资源，故采用参数两更小的XLM-RoBERTa-base网络；在网络搭建时，使用huggingface代码仓库开源的XLM-RoBERTa-base网络，使用其中的config.json，pytorch_model.bin，sentsncepiece.bpe.model，tokenizer_config.json文件作为初始网络结构及权重参数和初始分词器结构及参数，完成本次系统所需特征提取器网络及其分词器的构建。

2.2.3 模型训练

导入如pandas等相关库，同时，编写Config类集中管理模型路径和使用的超参数（最大长度128等）自动检测GPU加速。编写自定义的ViolenceDataset数据集类，使用pandas读取CSV数据时严格执行数据清洗。用XLM-RoBERTa分词器将文本转化为包含input_ids和attention_mask的张量。训练过程中，每个epoch先以AdamW优化器进行梯度更新，计算平均训练损失，随后在验证集上评估准确率和F1值，动态保存当前epoch模型和最佳F1模型。代码特别注重异常处理和资源优化，通过分批次处理和PyTorch的DataLoader实现内存高效利用，同时输出每个epoch耗时以便监控训练效率，在整体流程的安排上兼顾模型性能与训练的稳定性。

2.2.4 模型小测

经过200个epoch的训练之后，模型测试结果汇总如左图3.6，可以看到模型的类别分布均匀，占比接近1：1；此外，模型对于两类类别预测的准确率均高于0.91，f1值均高于0.9，召回率为0.91，效果很好，符合预期。

2.3 次级分类网络构建

2.3.1 网络改进

考虑到用于次级网络的训练的数据集仅为COLD数据集，且总共只有3类：race，region，gender。故考虑冻结xlm-roberta-base网络本体的预训练参数，使其不参与训练，因为预训练参数已经能很好的提取出输入文本的语义；但考虑到经网络提取出的语义向量仍然比较抽象，单层分类头不能很好的拟合，故采用自定义分类头参与训练，自定义分类头的结构如图4.7所示，添加两层隐藏层dense1、dense2用于更好的拟合，其神经元个数分别为96、48，总的参数量只有98k，相较于两层分别为512、256个神经元或一层隐藏层256个神经元大幅降低了参数量，有效的缩短了训练一轮所消耗的时间。其次，在每一层隐藏层后都添加了一层层标准化（Layer Normalization），其核心作用是稳定训练动态，能对每个样本的特征维度进行标准化（均值0、方差1），缓解内

部协变量偏移；并且还能使激活值分布更稳定，且允许使用更大的学习率，加速收敛，同时还能轻微的抑制过拟合，尤其在小数据集上效果显著。不仅如此，我还在其中添加了残差连接（Residual Connection），它的核心作用是解决梯度消失问题，在深层网络中，梯度反向传播时可能逐层衰减，残差连接通过“跳跃连接”（Shortcut）直接将输入传递到深层，保持梯度流动。而且残差连接还能促进特征的复用，允许网络直接学习输入与输出的差异（残差），而非完整的映射，降低学习难度。其次，添加残差链接在遇到网络层数增加时，也能保持训练的效率，从而起到一个稳定训练的作用。

在激活函数的选择方面，选择RELU作为激活函数，目的是计算高效，有效缓解训练过程中可能会遇到的梯度消失问题。

在分类头中添加dropout，避免网络学到过多的噪声从而过拟合。

2.3.2模型训练

预训练的XLM-RoBERTa-base模型构建特征提取网络，冻结预训练层参数的方式保留其跨语言特征提取能力，自定义一个包含RELU激活函数、层归一化、残杀链接的、两隐藏层的自定义分类器，神经元数量分别设定为96和48。训练过程中采用动态学习率策略，将AdamW优化器的学习率设置为1e-4，仅对分类器参数进行更新。用标准化的批次处理方式，每个epoch包含完整的训练集前向传播、损失回传（反向传播）和验证集评估环节。损失函数直接调用模型内置的交叉熵计算模块，配合每批次梯度清零和参数更新机制，确保整个优化过程的准确。验证阶段，设置同步计算准确率和F1值双指标，采用基于GPU加速的并行计算策略，在500个训练周期内持续监测模型性能。该模块伪代码同主分类网络模块在网络的结构构建上不同，其他流程上保持相同。

2.4用户交互界面实现

在用户交互界面的搭建上使用tkinter作为搭建工具。

用户交互界面搭建效果：用户可将需要检测的文本输入进“输入文本”框内，点击“检测”按钮即可开始检测，检测后的结果会在“检测结果”框内出现。

3 系统测试

3-1存在语言暴力且次级分类为race

3-2存在语言暴力且次级分类为region

3-3存在语言暴力且次级分类为gender

3-4不存在语言暴力

结论

本设计内容为基于Transformer的语言暴力检测系统，系统采用xlm-roberta-base网络作为语言暴力的主要与次级分类器（特征提取器），使用tkinter搭建用户交互界面，经过环境搭建，收集数据集、数据集清洗、数据集各项预处理、网络搭建、网络训练、测试等各项流程，取得了不错的结果，能实现系统所有功能的正常运行。

致谢

行文至此，意味着我的本科生涯即将画上句点。回首求学路上的点滴，心中充满感激之情。在此，谨向所有给予我支持与帮助的老师、同窗、亲友致以最诚挚的谢意。感恩我的父母和家人。二十余载求学路，是你们始终如一的理解与支持让我心无旁骛地追逐学术理想。特别感谢我的父母，你们的包容与鼓励始终是我最坚实的后盾。

山水一程，终须一别。谨以此文致敬我的青春岁月，愿初心不改，步履不停。

参考文献

[1]Kotsakis R, Vrysis L, Vryzas N, et al. A web framework for information aggregation and management of multilingual hate speech[J]. Heliyon, 2023, 9(5): e16084.

[2]Mozafari M, Farahbakhsh R, Crespi N. Cross-lingual few-shot hate speech and offensive language detection using meta learning[J]. IEEE Access, 2022, 10: 14880-14890.

[3]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.

[4]Bouchene M M, Abainia K. Classical machine learning and transformer models for offensive and abusive language classification on Dziri language[C]//2023 International Conference on Decision Aid Sciences and Applications (DASA). Annaba, Algeria: IEEE, 2023: 116-120.