这是发表于 CCIS 第 1798 卷的书籍章节预印本，可通过以下链接访问：
https://doi.org/10.1007/978-

达罗毗荼语系语言中的恐同与恐跨检测：探索深度学习方法

Deepawali Sharma、Vedika Gupta 和 Vivek Kumar Singh

摘要。在线社交媒体平台上辱骂性内容的增加，正在影响在线用户的社会生活。冒犯性言论和仇恨言论的使用使社交媒体变得有害。恐同和恐跨构成了针对 LGBT+ 群体的冒犯性评论。检测并处理这些评论至关重要，以便及时标记相关内容，或向从事此类行为的用户发出警告。然而，自动检测此类内容是一项具有挑战性的任务，在被视为低资源语言的达罗毗荼语系语言中尤其如此。受此启发，本文尝试探索不同深度学习模型在马拉雅拉姆语和泰米尔语社交媒体评论分类任务中的适用性，将其分为恐同、恐跨和非反LGBT+内容。本文将常用的深度学习模型——卷积神经网络（CNN）、使用 GloVe 嵌入的长短期记忆网络（LSTM），以及基于Transformer的学习模型（多语言BERT 和 IndicBERT）应用于该分类任务。结果表明，IndicBERT 的表现优于其他已实现的模型；在马拉雅拉姆语和泰米尔语上的加权平均F1分数分别为 0.86 和 0.77。因此，本研究证实，IndicBERT 在所选达罗毗荼语系语言的该任务上表现更佳。

关键词：深度学习、恐同、马拉雅拉姆语、泰米尔语、恐跨。

1 引言

人们会在各种社交媒体平台上分享和发布自己对各种话题的观点、看法和想法。创作此类内容的自由也导致互联网上出现大量令人不快的内容，涉及种族主义、恐同、恐跨，以及针对特定社群和组织的攻击。不幸的是，LGBT+ 人群会遭到其他社群在网上辱骂，进而出现心理障碍（焦虑障碍、行为和情绪障碍、抑郁症）。有时，这些弱势社群上线寻求支持时，会有人发起针对他们的行动，而这些行动往往会演变为欺凌[8]。自动检测此类内容是一项艰巨的任务；如果文本使用的是资源匮乏语言，难度更大。许多研究已针对某些特定语言中的仇恨言论[9][10][11]，、冒犯性和辱骂性内容[12][13][14]，、网络欺凌[15][16][17]等内容开展检测工作。然而，对于资源匮乏的印度语言网络文本中的种族主义、恐同和恐跨内容，相关检测工作仍然非常有限。

本文尝试朝这一方向开展研究。研究在泰米尔语和马拉雅拉姆语两种语言的数据上实现了多种基于深度学习的模型（卷积神经网络、长短期记忆网络、多语言BERT 和 IndicBERT）。这两种语言都属于达罗毗荼语系语言，且资源匮乏。实验使用了 Dra- vidianLangTech [8] 提供的数据集。评论被分类为恐同、恐跨和非反 LGBT+内容。包含针对男同性恋者、女同性恋者或其他社群的贬损性标签或侮辱性短语的评论被归类为恐同[8]。可能是同性恋者，或可能表现出恐跨但不恐同的人，被识别为恐跨者。因此，宣扬或提议任何不利于跨性别者的内容的评论被归类为恐跨[8]。不包含恐同或恐跨内容，且对读者有积极影响的评论，则被归类为非反 LGBT+内容。

本文采用不同的基于深度学习的模型，检测评论中的恐同和恐跨现象。具体而言，本文完成了以下工作：

- 使用 GloVe 词嵌入，为两种语言（马拉雅拉姆语和泰米尔语）的数据实现深度学习模型（卷积神经网络、长短期记忆网络）。
- 为两种语言（马拉雅拉姆语和泰米尔语）的数据实现基于 Transformer 的模型（多语言BERT、IndicBERT）
- 使用标准指标，对在标准数据集上实现的不同模型进行性能评估。

本文其余部分安排如下：第 2 节讨论恐同和恐跨检测的相关工作。第 3 节介绍数据集及其维度。第 4 节介绍实验设置。第 5 节展示结果。第 6 节简要总结结果及得出的主要结论，并以此结束全文。

2 相关工作

关于检测社交媒体评论中恐同和恐跨内容的既有研究非常有限，尤其是在低资源印度语言方面。一项研究采用集成式 Transformer 模型，对泰米尔语和泰米尔语-英语数据集中的社交媒体评论进行恐同和恐跨分类 [1]。类似地，为了将英语评论分类为恐同和恐跨，研究者采用了不同的基于 Transformer 的模型（BERT、RoBERTa 和 HateBERT）[2]。为检测恐同和恐跨，研究者在英语、泰米尔语和泰米尔语-英语数据集上试验了不同的单语和多语基于 Transformer 的模型 [3]。另一项研究针对英语、泰米尔语和英语-泰米尔语语码混合数据集，试验了基于 Transformer 的模型（RoBERT-base）[6]。一项研究在泰米尔语、英语和泰米尔语-英语中采用句子嵌入和集成模型实现了神经网络 [4]。一些研究利用词频-逆文档频率，采用不同的分类算法，将英语、泰米尔语和泰米尔语-英语中的 YouTube 评论分类为恐同和恐跨 [5]。类似地，词嵌入与支持向量机（支持向量机）相结合，并与 BERT-

基于的模型一同应用于英语、泰米尔语和英语-泰米尔语 [7]。然而，此前尚无针对马拉雅拉姆语社交媒体评论中恐同和恐跨检测的研究。此外，既有研究尚未探索诸如 IndicBERT 等较新的方法。因此，本研究探讨了包括近期提出的 IndicBERT 模型在内的各种基于深度学习的模型，是否适用于检测马拉雅拉姆语泰米尔语言社交媒体评论中的恐同和恐跨任务。

3 数据集描述

本研究使用的数据集由 DravidianLangTech 发布，可在线访问：https://codalab.lisn.upsaclay.fr/competitions/。该数据集包含两种达罗毗荼语系语言（马拉雅拉姆语和泰米尔语）的评论。数据集包含两个属性：评论和类别。评论属性以文本形式呈现人们的情感/观点，类别属性则说明评论属于以下三类中的哪一类：恐同、恐跨和非反LGBT+内容。马拉雅拉姆语数据集包含 3,114 条评论，泰米尔语数据集包含 2,662 条评论。表 1 展示了马拉雅拉姆语数据集中各类别评论的分布情况。表 2 展示了泰米尔语数据集中各类别评论的分布情况。

表 1. 马拉雅拉姆语数据集中评论的类别分布

类别	评论数
恐同	2434
恐跨	491
非反LGBT+内容	189

表 2. 泰米尔语数据集中评论的类别分布

类别	评论数
恐同	2022
恐跨	485
非反LGBT+内容	155

3.1 数据预处理

由于数据预处理是自然语言处理应用和模型构建中的重要步骤，我们对原始文本进行了以下预处理：

- 移除标点符号• 移除数字• 删除评论中的网址。• 删除不包含任何额外信息的重复字符。

- 删除预处理过程中产生的多余空格，或评论中原有的多余空格。

4 实验设置

4.1 卷积神经网络

卷积神经网络最初用于图像处理，但近年来也在自然语言处理任务中取得了良好效果。将卷积神经网络应用于文本而非图像时，会使用一维卷积神经网络。使用一维卷积神经网络需要词嵌入层。本文使用 GLoVe 词嵌入。为模型准备词嵌入时，首先移除所有标点符号、链接等，将文本分词，再使用 pad_sequences() 对文本进行填充，使所有序列长度相同。图 1 展示了卷积神经网络的实现过程。通过词嵌入，词被表示为向量。嵌入矩阵被传入嵌入层。之后，添加卷积层，并使用“relu”作为激活函数。接着添加池化层，以降低维度的复杂性；具体而言，使用最大池化。最后添加全连接层，也称为稠密层，并使用“Softmax”作为激活函数，因为该数据集采用多分类。因此，模型将评论分为三类：恐同、恐跨和非反 LGBT+内容。该模型针对两种语言（马拉雅拉姆语和泰米尔语）各训练 100 轮，并使用“Adam”作为优化器。用于对推文分类的损失函数是“分类交叉熵”。

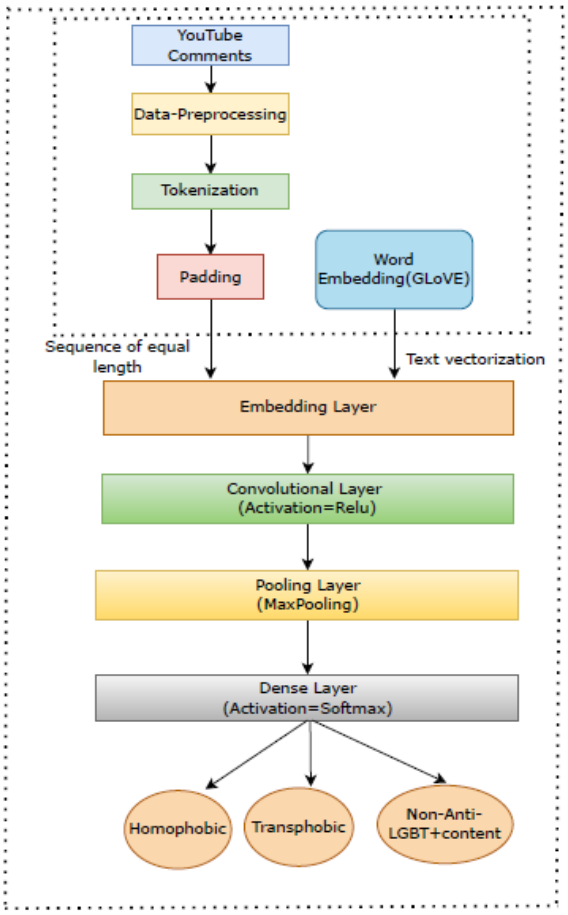

图 1. 使用卷积神经网络检测恐同和恐跨

4.2 长短期记忆网络

长短期记忆网络用于文本分类，因为它能够有效记忆信息并剔除无关信息。使用适当的嵌入层后，长短期记忆网络模型能够理解输入字符串的含义，并将文本分类到最精确的输出类别。在添加嵌入层之前，首先对数据进行预处理，并将文本分词为词元。之后，在序列末尾补零，使所有序列长度相同。接着，使用词嵌入（GLoVe）将词表示为向量。填充后的序列被作为输入传入嵌入层。在嵌入层之后添加长短期记忆层，并使用 ReLU 作为激活函数。

最后添加一个以 Softmax 为激活函数的全连接层，将评论分类为两个以上的类别。图 2 展示了用于评论分类的 LSTM 模型框图。该模型针对两种语言均使用分类交叉熵作为损失函数、使用“Adam”作为优化器，训练 100 轮。

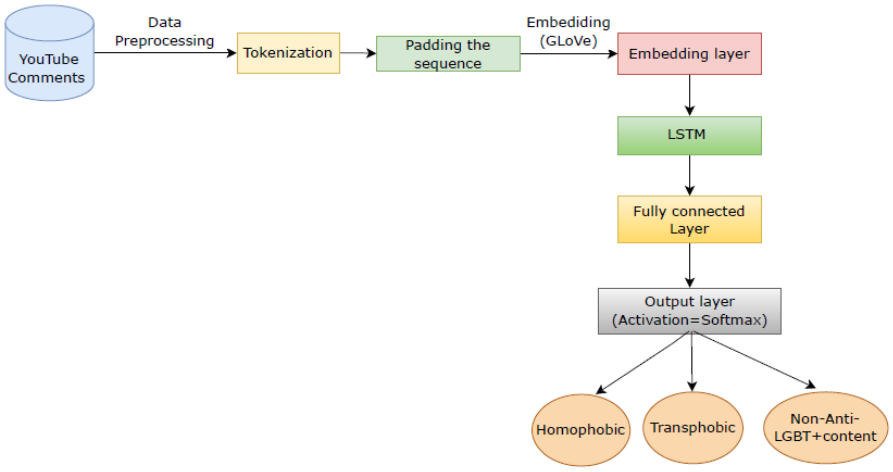

图 2. LSTM 模型实现框图

4.3 多语言BERT

BERT 于 2018年由 Jacob Devlin 及其谷歌同事开发。BERT 使用大规模英语语料进行训练，而多语言BERT则使用包括英语在内的 104 种不同语言进行训练[19]。为每种语言分别建立 BERT 模型并不可行，因此引入多语言BERT，作为适用于 104 种语言的单一 BERT 模型。本文使用的是 BERT 基础版，它包含 12 层 Transformer 编码器。首先，借助 BERT分词器将评论转换为词元。图 3 展示了 BERT 对评论进行分类的框架。序列的开头和结尾分别添加两个特殊词元 [CLS] 和 [SEP] 。每个词元对应的词汇表ID即为词元嵌入。句子嵌入用于区分不同句子；位置嵌入则用于表示序列中每个词的位置。该模型基于预训练的BERT基础版多语言大小写区分模型构建。批量大小为 32，学习率为 3e- 5，训练轮次=5 用于在数据集上训练模型。最后，分类器将评论分为恐同、恐跨和非反 LGBT+内容类别。

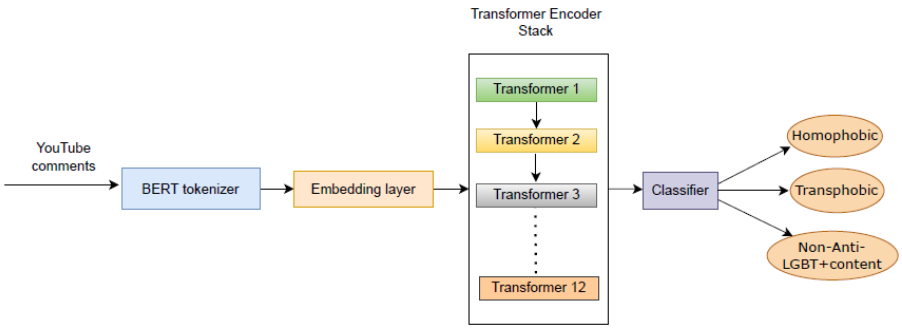

图 3. 基于 Transformer 的 BERT 模型框架

4.4 IndicBERT

IndicBERT 是一个使用 12 种印度语言训练的多语言模型：泰米尔语、马拉雅拉姆语、孟加拉语、马拉地语、旁遮普语、英语、印地语、古吉拉特语、阿萨姆语、奥里亚语、泰卢固语和卡纳达语。与多语言BERT 和 XLM-R 等其他模型相比，它的参数数量较少，但在各种任务上仍取得了最先进（SOTA）的性能[18]。IndicBERT 基于 ALBERT 模型，而 ALBERT 是 BERT 的衍生模型。文本使用 Autotokenizer 进行分词，并像 BERT 模型一样，在序列开头添加 [CLS] 词元，在序列末尾添加 [SEP] 词元。IndicBERT 还能够将多个句子作为单个序列输入。由于该数据集使用印度语言：马拉雅拉姆语和泰米尔语，因此在此数据集上，IndicBERT 的表现优于其他模型；其在马拉雅拉姆语和泰米尔语上的加权平均 F1 分数分别为 0.86 和 0.77。参数经过微调：学习率设为 3e-5，训练轮次=5 ，批量大小为 32。最后，模型将评论分为三个类别：恐同、恐跨和非反 LGBT+内容。

5 结果

我们使用卷积神经网络和基于 Transformer 的模型，将两种语言的评论分类为以下三类之一：“恐同”“恐跨”和“非反 LGBT+内容”。本节展示所实现模型的性能。表 3 展示了用于马拉雅拉姆语的卷积神经网络、长短期记忆网络和迁移学习模型（多语言BERT 和 IndicBERT）的性能分析。结果显示，卷积神经网络、长短期记忆网络、多语言BERT 和 IndicBERT 的加权平均 F1 分数分别为 0.80 、 0.81 、 0.83 和 0.86。对于马拉雅拉姆语，IndicBERT 的表现优于其他模型。表 4 展示了所实现模型在泰米尔语上的性能分析。深度学习模型（卷积神经网络和长短期记忆网络）的加权平均F1分数分别为 0.71

和 0.72。基于 Transformer 的模型（IndicBERT）的表现优于其他三个已实现的模型（卷积神经网络、长短期记忆网络和多语言BERT），加权 F1 分数为 0.77。

表 3. 马拉雅拉姆语数据集上所实现模型的性能分析

模型	恐同			恐跨			非反 LGBT+内容			加权平均 F1-分数
	P	**R**	**F1-分数**	**P**	**R**	**F1-分数**	**P**	**R**	**F1-分数**	
卷积神经网络(GloVe)	0.65	0.35	0.45	0.86	0.32	0.46	0.84	0.96	0.89	0.80
长短期记忆网络(GloVe)	0.69	0.45	0.54	0.60	0.32	0.41	0.86	0.95	0.90	0.81
多语言BERT	0.74	0.47	0.58	0.67	0.32	0.43	0.86	0.96	0.91	0.83
IndicBERT	0.79	0.49	0.59	0.70	0.39	0.50	0.88	0.97	0.91	**0.86**

*P=精确率，R=召回率

表 4. 泰米尔语数据集上所实现模型的性能分析

模型	恐同			恐跨			非反LGBT+内容			加权平均 F1-分数
	P	**R**	**F1-分数**	**P**	**R**	**F1-分数**	**P**	**R**	**F1-分数**	
卷积神经网络（GloVe）	0.57	0.08	0.14	1.00	0.27	0.42	0.78	0.99	0.88	0.71
长短期记忆网络(GloVe)	0.67	0.12	0.21	0.80	0.27	0.40	0.79	0.98	0.87	0.72
多语言BERT	0.70	0.15	0.24	0.79	0.30	0.43	0.81	0.96	0.87	0.74
IndicBERT	0.72	0.18	0.28	0.76	0.34	0.46	0.84	0.94	0.88	**0.77**

*P=精确率，R=召回率

6 结论

本文分析了不同深度学习模型（卷积神经网络、长短期记忆网络、多语言 BERT 和 IndicBERT）在检测马拉雅拉姆语和泰米尔语文本中的恐同和恐跨内容这一任务上的表现。研究目标是将每条评论归入给定的三类之一：恐同、恐跨和非反LGBT+内容。实验结果显示，对于马拉雅拉姆语和泰米尔语，IndicBERT模型均优于其他模型，其报告的加权平均F1分数分别为0.86和0.77。尽管其他模型也表现尚可，但就测得的性能值而言，IndicBERT模型仍是所有模型中表现最好的。因此，在所选达罗毗荼语系语言中，针对恐同和恐跨检测任务，IndicBERT模型成为各种所提方法中的佼佼者。当前研究还可扩展至其他印度语言的数据。此外，也可以探索将基于集成的模型用于该任务。

参考文献

1. Upadhyay, I. S., Srivatsa, K. A., Mamidi, R.：Sammaan@ lt-edi-acl2022：用于检[illegible]言技[illegible]LT-I[illegible]容语[illegible] Bhandari, V., Goyal, P.：bits[illegible]容。[illegible]Cap[illegible]语中[illegible]140-[illegible]Nay[illegible]cl2022：使用支持向量机检测平等、多样性与包容领域中的恐同/恐跨内容[illegible]202[illegible]RoB[illegible]测方法。载于：第二届平等、多样性与包容语言技术研讨[illegible]Gay[illegible]和基[illegible]ransformer模型检测多种语言中的恐同/恐跨内容。载于：第二届平等、多样[illegible]Cha[illegible]Sam[illegible]恐跨[illegible]S. Kamal, A. Fazil, M., [illegible]使用[illegible]《IEEE Access》, 10, [illegible]A., A[illegible]意力[illegible]志》[illegible]Smi[illegible]集, [illegible]78页（2019年7月）。12. Koufakou, A., Pamungkas, E. W., Basile, V., P[illegible]届网[illegible]43页。计算语言学协会（2020年）。13. Susa[illegible]测冒[illegible]IEEE[illegible]潜在[illegible]性语[illegible]

15. Cheng, L., Li, J., Silva, Y. N., Hall, D. L., Liu, H.：Xbully：在多模态语境下进行网络欺
（2(lez, T., Arabnia, H. R.：利
用大
252- ）。17. Hani, J., Mohamed, N., Ahmed, M., Emad, Z., Amer,
E., A
应用 ikwani, D., Kunchukuttan, A., Gokul, S.,
Gok M., M., Kumar, P.：IndicNLPSuite：印度
语言
EMI
K.,T 本.
arXi