用于检测威胁生命文本的大型语言模型

Thanh Thi Nguyen*, Campbell Wilson, and Janis Dalins

AiLECS Lab, Faculty of IT, Monash University, Melbourne, VIC 3800, Australia { thanh.nguyen9,campbell.wilson,janis.dalins } @monash.edu

Abstract. 检测危及生命的语言对于保护陷入困境的个人，促进心理健康和福祉，以及防止潜在的伤害和生命损失至关重要。本文提出了一种有效的方法，通过使用大型语言模型（LLMs）识别危及生命的文本，并将其与传统方法进行比较，如词袋、词嵌入、主题建模和双向编码器表示的转换器。我们微调了包括 Gemma、Mistral 和 Llama-2 在内的三个开源 LLMs，使用其 7B 参数变体在不同的数据集上进行训练，这些数据集包括类平衡、失衡和极端失衡场景。实验结果表明，LLMs 相比于传统方法表现出强劲的性能。具体来说，Mistral 和 Llama-2 模型在平衡和失衡数据场景中表现优秀，而 Gemma 略逊一筹。我们采用了上采样技术来处理失衡数据场景，并证明虽然这种方法对传统方法有好处，但对 LLMs 的影响不大。这项研究展示了 LLMs 在现实世界中识别危及生命的语言问题上的巨大潜力。

1 引言

鉴别威胁生命的语言至关重要，原因有几个。检测到表示自我伤害或伤害他人的语言使干预成为可能，以在发生任何伤害之前进行干预。这可能涉及到警告执法部门、联系急救服务，或为处于危机中的个人提供支持。机器学习方法常用于识别仇恨言论或攻击性内容等有害语言。然而，它们在检测威胁生命的语言方面的应用相对罕见，尤其是在英语中。我们在文献中仅发现了少数此类研究。例如，在 [11] 中引入了一种用于自动检测威胁性乌尔都文本的方法。其性能通过使用叠加分类模型和特征提取技术得到了提升。同样，[1] 中的研究提出了两个共享任务，专注于检测乌尔都语中的辱骂和威胁性语言。两个任务都被框定为二元分类挑战，系统负责将乌尔都语推文分类为：第一个任务为辱骂和非辱骂，第二个任务为威胁和非威胁。

最近，[12] 的工作设计了一个多语言文本分类框架，以解决在英语和乌尔都语中检测威胁内容的挑战。然而，与许多传统文本分类方法类似，他们的方法涉及一个文本预处理阶段，包括删除标点符号、话题标签、数字、HTML 标签、URL，替换表情符号和颜文字，解决拼写错误问题，以及解码英文缩写等多个步骤。那些预处理步骤既繁琐，又具有主观性，并且在不同的数据集中表现不一致。这可能会显著影响分类模型在不同未见测试数据集上的性能和一致性。

在这项研究中，我们提出了一种使用开源大语言模型（LLMs）如 Gemma、Mistral 和 Llama-2 进行威胁生命文本检测的方法。通过利用 LLMs 的强大功能，我们的方法对噪声数据具有抗性，而不需要像现有工作那样进行文本预处理步骤，如词形还原、去除标点符号、去掉标签、多余的空格、停用词等。我们

* Corresponding author.

的基于 LLM 的方法避免了特征提取和分类的单独设计，这通常可能涉及大量的试验和错误来找到这些步骤的最佳组合。我们展示了基于 LLM 的方法在不同数据场景下的一贯性能，包括平衡、不平衡和极度不平衡的数据集。我们对我们的方法与传统方法（TF-IDF、词嵌入、主题建模和 BERT）在六个数据集上的比较进行了全面分析，并强调在评估具有类别不平衡的分类方法时，不仅要使用准确率，还要使用 F 分数和 AUC 的重要性。

2 竞争方法

2.1 TF-IDF

TF-IDF 是一种用于在词袋模型中对术语进行加权的方法，该模型将文本数据表示为一组单词，忽略语法和单词顺序。虽然 TF-IDF 在向量空间中提供了文本数据的表示，但它并没有像词嵌入方法那样固有地获取单词之间的语义关系。

方法如 GloVe 和 Word2Vec，通过基于共现模式创建词嵌入来学习语义关系，其中相似的词具有相近的表示。

GloVe GloVe，即用于词表示的全局向量的缩写，通过创建一个共现矩阵来运作，该矩阵包含每个词在其他词的上下文窗口中出现的频次计数。

Word2Vec 这个范式包含两种主要方法：连续词袋（CBOW）和跳字模型。CBOW 使用周围的上下文词来预测目标词。跳字模型则相反，从给定的目标词预测上下文词。

2.2 主题建模

这是一种在自然语言处理中用于发现文档集中的潜在主题或主题的无监督学习技术。常见的主题建模方法包括潜在狄利克雷分布（LDA）和潜在语义索引（LSI）。LDA 假设每个文档是主题的混合体，每个词被分配给一个主题，通过迭代调整主题分布以最大化给定文档的可能性。LSI 则使用对项-文档矩阵进行奇异值分解进行降维，揭示潜在的语义结构并构建索引以提高文档检索效率。

2.3 BERT

BERT，代表来自变压器的双向编码器表征，是一个预训练的自然语言处理模型，由 [2] 介绍。与之前仅在一个方向上（左到右或右到左）顺序处理文本数据的语言处理方法不同，BERT 是双向的，这意味着它可以同时考虑词语左侧和右侧的全上下文。这有助于获得更精确的词语和短语表示。通过在特定的文本分类任务上微调 BERT，它可以在各个领域中取得卓越的表现。在本研究中，我们使用 BERT-en-uncased 模型，其编码器和预处理器分别从 [15] 和 [16] 获取。

2.4 LLM - Gemma 7B

Gemma 是由 Google DeepMind 提出的，描述在 [4] 中，并提供两种变体：一种是具有 70 亿参数的模型，专为在 GPU 和 TPU 上的高效开发和部署而定制，另一种是具有 20 亿参数的模型，优化用于 CPU 和设备上的应用 [4] 。

7B 模型采用多头注意力机制，而 2B 模型使用多查询注意力。Gemma 模型使用了来自 Gemini 的 SentencePiece 分词器 [8] 的一个子集。这个分词器对数字进行拆分，保留额外的空白，并使用字节级编码来处理未知符号。每层中使用旋转位置嵌入 [14] 代替绝对位置嵌入。此外，为了减少模型尺寸，输入和输出之间的嵌入是共享的 [4] 。

2.5 LLM - Mistral 7B

Mistral 7B LLMs 由预训练的 Mistral-7B-v0.1 模型和几个 Mistral 7B Instruct 变体 [7] 组成。这些模型配备了 70 亿个参数，利用分组查询注意力 (GQA) 和滑动窗口注意力 (SWA)。GQA 明显提升了推理速度并降低了解码时的内存需求，使得可以加大批处理大小，从而提高吞吐量。

SWA 被设计用来更高效地处理扩展序列，同时降低计算成本，从而解决使用 LLMs 时的一个常见限制。SWA 利用变压器的多层结构将其注意能力扩展到超出固定窗口大小 W 的范围。在每一层中，位置 i 的隐藏状态，记作 h_i ，会考虑前一层中所有位置落在 $i - W$ 到 i 范围内的隐藏状态。

Llama-2 系列在发布时包括预训练和微调的变体，每个变体提供了三种模型，分别具有 7B、13B 和 70B 可训练参数。这些预训练变体是通过利用包含两万亿个标记的庞大文本语料库的自监督学习方法获得的。随后，这些预训练变体通过监督微调程序和带有人类反馈的强化学习，采用指令数据集和专为对话使用案例定制的人类标注数据进行了进一步优化。

与绝对位置编码不同，Llama-2 使用旋转位置嵌入来编码标记的位置信息。对于分词，LLaMA 分词器使用源自 SentencePiece 的字节对编码 [8] 。因此，Llama 的词汇表包含 32k 个标记 [17] 。

3 使用 LoRA 微调 LLMs

像 Gemma、Mistral 和 Llama-2 这样的 LLMs 通过在大量文本数据上的预训练积累了丰富的世界知识。我们的方法是在这些模型之上进行微调，会根据每个微调数据集生成专门的专业知识，同时保留在预训练阶段获得的广泛的通用知识和推理能力。

我们使用低秩适配（LoRA）来微调大型语言模型（LLMs）的方法如图 1 所示。LoRA 是诸如 BitFit、SparseAdapter、AdaLoRA 等多种参数高效微调（PEFT）方法之一，可用于微调 LLM [9] 。LLM 的原始权重被转换为 Hugging Face 的 transformer 格式，以利用其提供的微调工具。这些转换后的模型随后被集成到 PEFT-LoRA 框架中，该框架基于 PEFT 库采用 LoRA [10] 。LoRA 方法通过汲取结构感知本征维度技术的灵感进行低秩微调。一个预训练权重矩阵 $M_0 \in \mathbb{R}^{d \times k}$ 的参数更新由两个低秩矩阵 M_A 和 M_B 的乘积决定：

$$\Delta M = M_A M_B \tag{1}$$

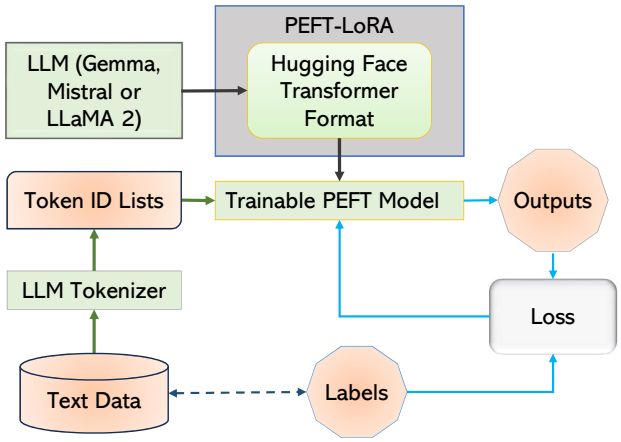

Fig. 1. 对预训练的大型语言模型进行微调以用于文本分类任务的方法。首先将文本数据通过大型语言模型的分词器转化为 token ID，然后将这些 token ID 作为输入，传入可学习的 PEFT 模型。该模型是通过将原始的大型语言模型权重转移至 Hugging Face 格式，并通过 PEFT-LoRA 机制加载构建的。

，其中 $M_A \in \mathbb{R}^{d\times r}$ 和 $M_B \in \mathbb{R}^{r\times k}$ 分别代表可学习参数的矩阵，其秩 r 显著小于 d 和 k 的最小值 [9] 。预训练模型的参数 M_0 保持不变，在训练期间不进行梯度更新。根据 Hu 等人的研究，M_0 和 $\Delta M = M_A M_B$ 都应用于相同的输入 x ；因此，$h = M_0 x$ 的前向传递的修改表达为：矩阵 M_A 以零初始化开始，而矩阵 M_B 以随机高斯值初始化。因此，在微调过程开始时，$\Delta M = M_A M_B$ 等于零。在整个训练过程中，$\Delta M x$ 按固定因子 $\frac{\alpha}{r}$ 进行缩放，其中 α 是一个常数 [5] 。在训练之后，可以通过将矩阵 $M_A M_B$ 相加，将可学习参数合并到原始权重矩阵 M_0 中。图 2 展示了 LoRA 及其与常规微调方法的区别。LoRA 使得小的可训练矩阵 M_A 和 M_B 可以进行训练，从而适应新的数据，同时最小化更新总数。通过绕过预训练权重的梯度计算，LoRA 方法显著减少了内存使用，使微调更加快速和高效。

我们的微调方法采用交叉熵损失函数，将神经网络的输出对数 $\hat{y}$ 与目标 y 进行比较。在图 1 中，“标签” 指的是目标 y ，而由可学习的 PEFT 模型生成的“输出” 表示对数 $\hat{y}$ 。交叉熵损失函数在输出 $\hat{y}$ 和目标 y 之间以小批量大小 N 取平均，表示为：

$$\mathcal{L}(\hat{y}, y) = \sum_{n=1}^{N} \frac{l_n}{\sum_{n=1}^{N} w_{y_n}} \tag{2}$$

，其中 w 表示一个权重向量，每个元素对应一个类别，而 l_n 定义为：

$$l_n = -w_{y_n} \log \frac{\exp(\hat{y}_{n,y_n})}{\sum_{c=1}^{C} \exp(\hat{y}_{n,c})} \tag{3}$$

，其中对数 $\hat{y}$ 是每个类别的未归一化对数，目标 y 包含类别索引，C 是类别总数。

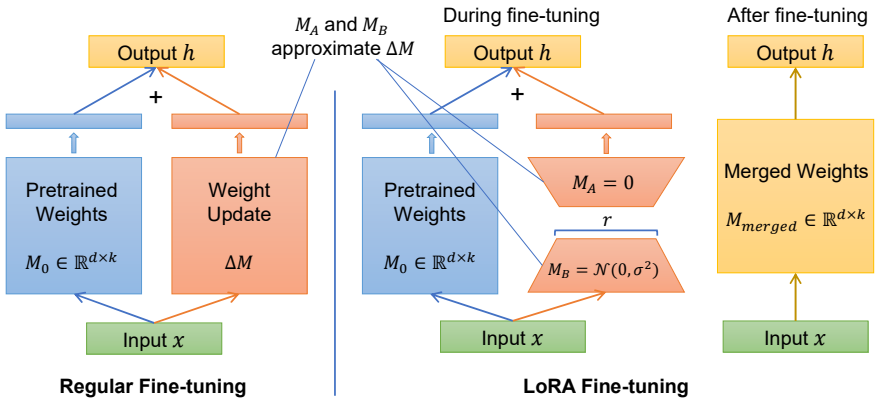

Fig. 2. 常规微调和 LoRA 微调之间的区别。LoRA 矩阵 M_A 和 M_B 近似于权重更新矩阵 ΔM ，其中内部维度的秩 r 是一个超参数。

Table 1. 六个实验数据集的组成

Dataset	No. of Threatening	No. of Non-Threatening	Overall Characteristic
1	915 texts in [3] and [18]	1,000 not-hate texts in [18]	Balance
2	915 texts in [3] and [18]	1,000 not-hate texts in [6]	Balance
3	915 texts in [3] and [18]	5,000 not-hate texts in [18]	Imbalance
4	915 texts in [3] and [18]	5,000 not-hate texts in [6]	Imbalance
5	915 texts in [3] and [18]	10,000 not-hate texts in [18]	Extreme Imbalance
6	915 texts in [3] and [18]	10,000 not-hate texts in [6]	Extreme Imbalance

本研究中使用的超参数设置如下。句子的最大长度设置为 128 个标记，较长的句子会被截断，较短的句子则会被填充，以确保大小统一。LoRA 的秩设为 8，训练的轮数为 10。我们使用随机梯度下降的 AdamW 优化器进行微调，学习率为 2×10^{-5} ，epsilon 系数为 10^{-8} 。此外，训练期间的小批量大小固定为 16。

虽然在执法、医疗保健、应急服务、心理健康和安全机构中识别威胁性语言很重要，但由于其稀缺性，收集威胁性文本以训练机器学习模型具有挑战性。文献中有几个仇恨言论数据集，但大多数缺乏真正的威胁性文本，只有“动态生成的仇恨言论”数据集例外。该数据集包含 18,969 个非仇恨文本和 22,175 个仇恨文本。仇恨文本分为几个类别，如贬低、敌意、非人化和威胁，其中威胁性文本占 606 个样本。我们在实验中将这 606 个样本作为威胁性文本的一部分。威胁性文本的另一部分来自威胁性英语语言（TEL）语料库，其中包含 309 篇书面文本。总共，我们有 915 个威胁性文本组成了小类的数据。

主要类别包括无威胁的文本。为了评估各种方法在处理平衡、不平衡和极度不平衡数据集时的表现，我们为无威胁文本创建了六种场景。

前三种情景基于上述“动态生成仇恨言论”数据集 [18] 中的 18,969 非仇恨文本，其中我们从这些非仇恨文本中无替换地抽取 1,000、5,000 和 10,000 个样本。

其他三个关于非威胁性文本的场景基于 [13] 中描述的数据集的公开发布，该数据集包括 135,556 个样本 [6] 。主要结果变量是“仇恨言论得分”，它是一个连续的仇恨言论测量，其中得分小于 0.5 表示非仇恨言论，而这一部分数据有 86,508 个样本。我们从这些非仇恨文本中不重复抽样 1,000、5,000 和 10,000 个样本，并将其视为非威胁性文本。

表 1 中呈现了与上述六个场景相对应的六个实验数据集的摘要。每个数据集都会随机分为训练集（90 %）和测试集（10 %），用于实验。为了应对类不平衡问题，使用上采样技术，通过随机有放回地抽取威胁文本（即，少数类），使训练集中威胁文本的数量与非威胁文本（即，多数类）的数量相匹配。

每个竞争方法都根据多个指标进行评估，包括准确度，F_β ，其 β 等于 0.5（即，$F_{0.5}$ 得分）、1.0（F_1 得分）和 2.0（F_2 得分）：

$$F_\beta = (1+\beta^2) \cdot \frac{\mathrm{P} \cdot \mathrm{R}}{(\beta^2 \cdot \mathrm{P}) + \mathrm{R}} \tag{4}$$

其中 P 是精确率，R 是召回率，而 β 代表精确率和召回率的相对重要性。F_1 得分决定精确率和召回率的调和平均数，在这两个指标之间取得平衡。它在类别分布不平衡的情况下特别有用。$F_{0.5}$ 得分是 F_1 得分的一个变体，它使精确率高于召回率。它在精确率比召回率更重要的情况下很有用。F_2 得分是 F_1 得分的另一个变体，它考虑到召回率比精确率更高。我们也使用 ROC 曲线下面积（AUC），因为它可以量化机器学习模型区分正负类别的能力。在这项研究中，我们将所有指标都以百分比形式表示。

4 结果与讨论

与我们的基于 LLM 的方法不同，该方法作为一种端到端方法，而使用 TF-IDF、词嵌入和主题建模的方法则需要与分类器集成，以对从这些技术获得的文本特征进行分类。为此，我们使用了一个集成分类器，结合了基于软投票机制的逻辑回归、支持向量机和随机森林。

4.1 在平衡数据集上的结果

针对两个平衡数据集，数据集 1 和数据集 2 的结果在表 2 中报告。TF-IDF 在所有指标中均取得了高分，而 GloVe 嵌入表现中等，但分数低于 TF-IDF。CBOW 和 skip-gram 的表现相似，显示出不错的结果，但都被 TF-IDF 超越。LSI 在两个数据集上的所有指标中都优于 LDA，但两者的表现均低于 TF-IDF 和 BERT。Gemma 不如 BERT，而 Mistral 表现出色，在数据集 2 中超过了大多数其他方法。Llama-2 在数据集 1 中的所有指标上都取得了最高分，但在数据集 2 中略落后于 Mistral。

评估了竞争的方法，既在提升采样的情况下也在没有提升采样的情况下，其中提升采样被用于缓解类不平衡问题。提升采样略微改善了 TF-IDF，并显著增强了 GloVe 嵌入、CBOW、skip-gram、LDA 和 LSI 的 F -scores 和 AUC。BERT 和 Gemma 的性能相当，而 Mistral 在大多数指标上优于它们。Llama-2 表现出色，但在所有指标上都不如 Mistral。对于 Gemma、Mistral 和 Llama-2 来说，提升采样在准确度、F -scores 和 AUC 上没有带来太大改进。

不平衡数据集 4（表 4 ） 与不使用上采样的方法相比，传统方法在使用上采样时在 F -score 和 AUC 上表现出提升。例如，GloVe 嵌入方法的 F_1 -score 从 44.44 % 提高到 55.68 %，而其 AUC 从 64.55 % 提高到 72.94 %。相反，使用上采样并未显著改善 LLMs。具体来说，它仅对 Gemma、Mistral 和 Llama-2 的性能产生轻微变化。

Table 2. 关于平衡数据集的结果 - 数据集 1 和数据集 2

Competing Methods	Dataset 1					Dataset 2				
	Acc.	F_1	$F_{0.5}$	F_2	AUC	Acc.	F_1	$F_{0.5}$	F_2	AUC
TF-IDF	77.08	72.15	80.74	65.22	76.60	83.33	82.22	83.90	80.61	83.22
GloVe Embedding	69.27	59.31	71.43	50.71	68.57	74.48	70.66	75.84	66.14	74.14
Word2Vec - CBOW	71.88	68.24	72.32	64.59	71.59	69.79	69.15	68.71	69.59	69.79
Word2Vec - Skip-gram	71.35	68.57	71.26	66.08	71.15	70.83	68.18	70.59	65.93	70.64
Topic Modeling - LDA	67.19	62.72	66.75	59.15	66.88	71.35	68.57	71.26	66.08	71.15
Topic Modeling - LSI	72.40	69.36	72.64	66.37	72.16	80.73	79.10	81.59	76.75	80.56
BERT - en-uncased	79.17	77.53	79.68	75.49	79.02	83.33	80.00	88.64	72.89	82.89
LLM - Gemma - 7B	75.52	74.59	74.84	74.35	75.48	80.73	78.11	83.12	73.66	80.43
LLM - Mistral - 7B	92.19	91.62	93.82	89.52	92.07	95.83	95.60	96.88	94.36	95.76
LLM - Llama-2 - 7B	93.23	92.82	94.38	91.30	93.14	95.31	94.97	97.25	92.79	95.19

Table 3. 数据集 3 的结果 - 第一个不平衡数据集

Competing Methods	Performance Metrics				
	Acc.	F_1	$F_{0.5}$	F_2	AUC
TF-IDF	88.85 (89.02)	56.00 (56.95)	65.42 (66.15)	48.95 (50.00)	71.08 (71.62)
GloVe Embedding	86.82 (88.18)	32.76 (53.33)	51.35 (62.31)	24.05 (46.62)	59.81 (69.80)
Word2Vec - CBOW	84.63 (78.04)	4.21 (38.68)	9.90 (36.03)	2.67 (41.75)	51.08 (64.23)
Word2Vec - Skip-gram	85.64 (83.28)	17.48 (53.52)	33.83 (49.74)	11.78 (57.93)	54.74 (74.33)
Topic Modeling - LDA	85.14 (82.94)	13.73 (51.67)	27.13 (48.47)	9.19 (55.33)	53.56 (72.82)
Topic Modeling - LSI	86.99 (82.60)	39.37 (54.22)	54.59 (49.11)	30.79 (60.52)	62.54 (75.68)
BERT - en_uncased	89.36 (78.38)	59.35 (54.93)	67.45 (45.51)	53.00 (69.27)	73.13 (80.61)
LLM - Gemma - 7B	87.84 (85.64)	56.63 (54.55)	61.04 (54.37)	52.81 (54.72)	72.66 (73.11)
LLM - Mistral - 7B	95.10 (92.06)	83.80 (70.06)	85.81 (78.80)	81.88 (63.07)	89.22 (78.67)
LLM - Llama-2 - 7B	94.59 (94.09)	81.40 (80.66)	85.57 (82.02)	77.61 (79.35)	86.73 (87.74)

(*) Values in brackets correspond to cases where upsampling is used.

Mistral 和 Llama-2 在多个指标上表现稳定，无论是否进行上采样。它们都获得了最高的总体准确率，其中 Mistral 达到了 95.78 %，Llama-2 达到了 96.96 %。Mistral 在该数据集上获得了最佳 AUC，得分为 92.25 %。

4.2 在极度不平衡数据集上的结果

极度不平衡数据集 5（表格 5 ） 由于该数据集（以及数据集 6）的极度不平衡性，所有方法的准确率都超过了 91 % ，这使得 F -分数和 AUC 成为评估更实际的指标。

与 GloVe 嵌入相比，TF-IDF 具有更好的准确性、F 分数和 AUC。虽然上采样对 TF-IDF 没有提升，但它显著增强了 GloVe 嵌入在 F 分数和 AUC 方面的表现。

CBOW 和 skip-gram 表现不佳，F 分数极低。经过上采样后，它们的 F 分数和 AUC 显著提高，但准确率下降。LSI 在所有指标上都优于 LDA。BERT 表现与 LSI 相当，两者均超过 Word2Vec 方法。Gemma 的表现不如 Mistral。Llama-2 展现出卓越的性能，达到了最高的准确率、F 分数和 AUC。上采样未能提升 LLM 方法的性能，即 Gemma、Mistral 和 Llama-2。

无论使用何种指标，Llama-2 和 Mistral 在该数据集中都表现出色。通常，过采样可以提升或至少保持传统（非 LLM）方法的性能。然而，即使进行过采样，非 LLM 方法在性能上依然明显不如 Mistral 和 Llama-2。

Table 4. 数据集 4 的结果 - 第二个不平衡数据集

Competing Methods	Performance Metrics				
	Acc.	F_1	$F_{0.5}$	F_2	AUC
TF-IDF	92.91 (92.91)	72.37 (73.08)	83.59 (82.61)	63.81 (65.52)	79.17 (80.04)
GloVe Embedding	88.18 (86.82)	44.44 (55.68)	62.22 (57.65)	34.57 (53.85)	64.55 (72.94)
Word2Vec - CBOW	86.15 (80.91)	25.45 (42.05)	43.48 (40.92)	17.99 (43.25)	57.23 (65.93)
Word2Vec - Skip-gram	87.50 (85.47)	38.33 (56.12)	57.21 (54.46)	28.82 (57.89)	61.96 (74.76)
Topic Modeling - LDA	86.66 (83.78)	32.48 (55.14)	50.26 (51.13)	23.99 (59.84)	59.71 (75.51)
Topic Modeling - LSI	90.71 (90.71)	62.59 (72.08)	74.43 (69.74)	53.99 (74.58)	73.93 (84.87)
BERT - en_uncased	91.55 (85.47)	67.11 (63.56)	77.51 (56.39)	59.16 (72.82)	76.62 (83.51)
LLM - Gemma - 7B	91.55 (90.03)	71.26 (63.80)	74.34 (69.71)	68.43 (58.82)	81.43 (76.15)
LLM - Mistral - 7B	95.78 (96.79)	86.63 (89.02)	86.35 (93.22)	86.91 (85.18)	92.25 (91.10)
LLM - Llama-2 - 7B	96.96 (96.96)	89.77 (89.66)	92.94 (93.53)	86.81 (86.09)	92.07 (91.63)

(*) Values in brackets correspond to cases where upsampling is used.

Table 5. 数据集 5 的结果 - 第一个极度不平衡的数据集

Competing Methods	Performance Metrics				
	Acc.	F_1	$F_{0.5}$	F_2	AUC
TF-IDF	92.58 (92.58)	31.93 (30.77)	49.74 (49.18)	23.51 (22.39)	59.75 (59.27)
GloVe Embedding	92.03 (91.39)	21.62 (47.19)	37.74 (49.18)	15.15 (45.36)	56.12 (70.05)
Word2Vec - CBOW	91.48 (84.89)	6.06 (33.73)	13.51 (29.54)	3.91 (39.33)	51.53 (66.49)
Word2Vec - Skip-gram	91.76 (89.10)	15.09 (43.60)	28.78 (41.14)	10.23 (46.37)	54.06 (70.70)
Topic Modeling - LDA	91.94 (87.55)	13.73 (33.33)	28.46 (32.02)	9.04 (34.76)	53.68 (64.13)
Topic Modeling - LSI	92.49 (88.55)	28.07 (45.89)	46.78 (41.47)	20.05 (51.36)	58.27 (73.73)
BERT - en_uncased	92.49 (87.45)	25.45 (49.45)	45.16 (41.93)	17.72 (60.25)	57.32 (79.80)
LLM - Gemma - 7B	93.96 (93.22)	57.14 (54.32)	66.47 (60.61)	50.11 (49.22)	72.41 (72.00)
LLM - Mistral - 7B	97.34 (96.15)	83.80 (76.14)	87.01 (79.95)	80.82 (72.67)	89.02 (84.56)
LLM - Llama-2 - 7B	98.26 (97.99)	89.62 (88.17)	91.72 (89.32)	87.61 (87.05)	92.86 (92.71)

(*) Values in brackets correspond to cases where upsampling is used.

极度不平衡数据集 6（表 6 ） 使用上采样时，TF-IDF 的所有指标都有轻微下降。与数据集 5 类似，GloVe 嵌入的方法相比 TF-IDF 精度、F -得分和 AUC 较低。上采样显著改善了其 F -得分和 AUC，但精度没有改善。

Table 6. 数据集 6 的结果——第二个极度不平衡的数据集

Competing Methods	Performance Metrics				
	Acc.	F_1	$F_{0.5}$	F_2	AUC
TF-IDF	94.60 (94.14)	59.86 (52.94)	72.61 (69.50)	50.93 (42.76)	72.76 (68.70)
GloVe Embedding	92.77 (92.12)	28.83 (52.22)	50.31 (54.02)	20.20 (50.54)	58.42 (72.83)
Word2Vec - CBOW	91.76 (87.91)	13.46 (36.54)	26.72 (34.73)	9.00 (38.54)	53.58 (66.24)
Word2Vec - Skip-gram	92.12 (91.12)	20.37 (52.68)	37.41 (50.47)	13.99 (55.10)	55.69 (75.61)
Topic Modeling - LDA	92.40 (92.31)	27.83 (48.78)	45.71 (53.91)	20.00 (44.54)	58.22 (69.60)
Topic Modeling - LSI	93.32 (91.30)	42.52 (56.62)	60.54 (52.45)	32.77 (61.51)	63.96 (79.52)
BERT - en_uncased	93.77 (86.54)	46.03 (51.16)	66.21 (41.89)	35.28 (65.70)	65.16 (84.06)
LLM - Gemma - 7B	97.16 (95.33)	82.68 (67.52)	85.85 (77.26)	79.74 (59.95)	88.45 (77.44)
LLM - Mistral - 7B	98.26 (98.26)	89.62 (89.50)	91.72 (92.26)	87.61 (86.91)	92.86 (92.38)
LLM - Llama-2 - 7B	98.08 (98.44)	88.77 (90.50)	89.63 (93.97)	87.92 (87.28)	93.23 (92.48)

(*) Values in brackets correspond to cases where upsampling is used.

BERT 表现出中等性能，与 LSI 相当，但不如 TF-IDF。当使用上采样时，其准确性和 $F_{0.5}$ 下降，而 F_1 、F_2 和 AUC 则显著增加。

Gemma 的表现异常出色，超过了所有传统方法。然而，当应用上采样时，Gemma 的所有指标都有轻微下降。Mistral 和 Llama-2 均表现出色，具有非常

高的准确性、F 分数和 AUC 值，超过所有其他方法。上采样对这些方法的性能没有太大影响。

总之，Mistral 和 Llama-2 在与传统方法的比较中始终表现出色，表明它们在处理文本分类任务方面的有效性，这要归功于它们复杂的架构。这些 LLM 方法有效地解决了类别不平衡问题，消除了上采样的必要性，尽管上采样确实提高了大多数传统方法的性能。此外，结果强调了除了准确性之外，F -分数和 AUC 指标在评估存在类别不平衡的分类方法时的重要性。

5 结论与未来工作

本研究探讨了多种方法，包括传统方法和 LLMs，以区分威胁性文本和非威胁性文本。结果表明，LLMs，尤其是 Mistral 和 Llama-2，一贯优于传统方法，如词嵌入、主题建模和 BERT。为了解决类别不平衡问题，我们在训练数据上使用了上采样技术。上采样通常增强了传统方法检测小类样本（即威胁性文本）的能力，从而提高了 F 分数和 AUC 值。然而，即便使用上采样，传统方法的表现仍不及 LLMs。值得注意的是，LLMs 在处理不平衡数据时表现出卓越的效果，无需上采样，这归功于其复杂的架构。

鉴于威胁性文本的稀缺，特别是在资源匮乏的语言中，未来的努力可以集中在收集这些语言中的文本，并探索基于大型语言模型的多语言方法。该多语言模型将能够无缝地跨语言运作，而无需对每种语言的文本数据进行单独的大型语言模型微调。未来研究的另一方向涉及利用量化方法提升大型语言模型的可访问性。不同的方法如量化的低秩适应、剪枝和秩增加的低秩适应，以及通用预训练变压器的量化可以被探索，因为每种方法都有其自己的优势和劣势。

References

1. Amjad, M., Zhila, A., Sidorov, G., Labunets, A., Butt, S., Amjad, H., Vitman, O., Gelbukh, A.: Overview of abusive and threatening language detection in Urdu at FIRE 2021. CEUR Workshop Proceedings **3159**, 744–762 (2021)
2. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: The North American Chapter of the ACL: Human Language Technologies. pp. 4171–4186 (2019)
3. Gales, T., Nini, A., Symonds, E.: The threatening English language (TEL) corpus (Jul 2022), https://doi.org/10.5281/zenodo.6815671
4. Gemma Team, Google DeepMind: Gemma: Open models based on Gemini research and technology (Feb 2024), https://storage.googleapis.com/deepmind-media/gemma/gemma-report.pdf
5. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)
6. Hugging Face: Dataset card for Measuring Hate Speech (Jan 2023), https://huggingface.co/datasets/ucberkeley-dlab/measuring-hate-speech
7. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., Casas, D.d.l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al.: Mistral 7B. arXiv preprint arXiv:2310.06825 (2023)

8. Kudo, T., Richardson, J.: SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In: Conference on EMNLP. pp. 66–71 (2018)
9. Lialin, V., Deshpande, V., Rumshisky, A.: Scaling down to scale up: A guide to parameter-efficient fine-tuning. arXiv preprint arXiv:2303.15647 (2023)
10. Mangrulkar, S., Gugger, S., Debut, L.: PEFT: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft (2022)
11. Mehmood, A., Farooq, M.S., Naseem, A., Rustam, F., Villar, M.G., Rodríguez, C.L., Ashraf, I.: Threatening Urdu language detection from tweets using machine learning. Applied Sciences **12**(20), 10342 (2022)
12. Rehan, M., Malik, M.S.I., Jamjoom, M.M.: Fine-tuning transformer models using transfer learning for multilingual threatening text identification. IEEE Access **11**, 106503–106515 (2023)
13. Sachdeva, P., Barreto, R., Bacon, G., Sahn, A., Von Vacano, C., Kennedy, C.: The measuring hate speech corpus: Leveraging Rasch measurement theory for data perspectivism. In: The 1st Workshop on Perspectivist Approaches to NLP @Language Resources and Evaluation Conference (LREC). pp. 83–94 (2022)
14. Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., Liu, Y.: RoFormer: Enhanced transformer with rotary position embedding. Neurocomputing **568**, 127063 (2024)
15. TensorFlow: Implementation of the BERT encoder API - Encoder (Oct 2020), https://tfhub.dev/tensorflow/bert_en_uncased_L-12_H-768_A-12/4
16. TensorFlow: Implementation of the BERT encoder API - Preprocessor (Oct 2020), https://tfhub.dev/tensorflow/bert_en_uncased_preprocess/3
17. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
18. Vidgen, B., Thrush, T., Waseem, Z., Kiela, D.: Learning from the worst: Dynamically generated datasets to improve online hate detection. In: The 59th Annual Meeting of the ACL. pp. 1667–1682 (2021)