arXiv:2304.00636v1 [cs.CL] 2 Apr 2023

CERIST'22：使用基于 BERT 的模型对推文进行 COVID-19 相关假新闻检测与情感分析

Rabia Bounaama[1], *, †, MohammedEl Amine Abderrahim[2], †

[1]特莱姆森大学生物医学工程实验室，阿尔及利亚，[2]特莱姆森大学阿拉伯语自然语言处理实验室，阿尔及利亚，

摘要

本文介绍了我们"techno"团队参与 CERIST'22 共享任务的情况。我们使用了一个与 COVID-19 疫情相关的现有数据集"task1.c"。其中包含 4,128 条用于情感分析任务的推文，以及 8,661 条用于假新闻检测任务的推文。我们将自然语言处理工具与最著名的预训练语言模型 BERT（来自变换器的双向编码器表示）相结合。结果表明，预训练语言模型效果显著：我们在情感分析任务中取得了 0.93 的准确率，在假新闻检测任务中取得了 0.90 的准确率。

关键词

CERIST、BERT、虚假新闻检测、情感分析

1. 引言

COVID-19影响了每个人的日常生活。截至2022年10月，全球已报告超过6亿例病例[1]。在COVID-19造成的当前中断局面下，每个人都有责任遵守政府为防止病毒传播而采取的措施，包括远程学习、隔离、居家和保持社交距离。这些措施引发了社交媒体用户的大量意见。分析这些意见最有效的方法之一是情感分析（SA）。SA是一种分类过程，通过书面内容识别用户的观点和情绪 [1]。

社交媒体平台上的信息可能包含真实或虚假的内容。遗憾的是，虚假内容传播迅速 [2]。假新闻（ FN ）是指通过传统新闻媒体或社交媒体蓄意传播虚假信息 [2]。COVID-19 疫情造成的严峻局势，尤其是新增病例数和死亡人数，引起了人们更多关注。如果这些信息有误，可能会对人们的心理健康产生负面影响，例如抑郁症和焦虑障碍。因此，FN检测已成为一个亟待研究的领域，

CERIST'22：*CERIST*自然语言处理挑战赛*2022*，*2023*年*3*月*29*日，阿尔及利亚，阿尔及尔
*通讯作者。 rabea.bounaama@univ-tlemcen.dz （R. Bounaama）； mohammedelamine.abderrahim@univ-tlemcen.dz (M. E.A. Abderrahim)
©2022 本文版权归作者所有。允许依据知识共享署名 4.0 国际许可协议（CC BY 4.0）使用本文。
orkshop Proceedings(CEUR-WS.org) [1]https://www.worldometers.info/coronavirus/

研究人员群体。FN检测旨在预测新闻文章是否为假新闻 [3]。在此背景下，开展实践研究迫在眉睫。为此，CERIST提出了若干待解决的任务。我们的参与涉及观点挖掘和情感分析任务，具体而言是“task1.c：COVID-19背景下的阿拉伯语情感分析与虚假新闻检测”[2]。

为解决情感分析和FN检测任务，可以采用多种方法。其中，效果较好的热门方法之一是使用BERT的迁移学习方法。使用预训练的BERT模型，无需从头训练模型，即可针对特定任务对其进行微调。这包括使用挑战赛组织者提供的标注数据集训练模型。模型的偏差和权重会根据情感分析或FN检测任务的要求进行调整。训练完成后，模型用于预测未见过的测试数据集中的标签。

本文其余部分安排如下：第2节介绍有关COVID-19评论中情感分析和FN检测的相关工作。第3节描述实验。第4节讨论所得结果。最后，第5节总结全文。

2. 相关工作

相关工作分为两个部分：与COVID-19相关的阿拉伯语情感分析，以及与COVID-19相关的阿拉伯语假新闻检测。

已有多项研究致力于应对COVID-19疫情相关的情感分析，并采用了不同的方法和技术。在这些方法中，深度学习是一种常用工具，尤其适用于推文等短文本。 [4]的研究针对阿拉伯语情感分析，运用深度学习分析了一个包含416,292条关于COVID-19的标注推文的语料库。作者发现，“corona”和“virus”是其语料库中的主导词。他们采用了传统机器学习方法，如支持向量机（SVM）、新型预训练Transformer模型“arabert”以及集成模型。结果发现，负向情感占主导，其次是正向和中性情感。作者得出结论：所提出的集成模型优于SVM、深度学习模型和AraBERT，准确率达到0.90。

COVID-19 对全球各国政府的决策产生了重要影响。例如，沙特阿拉伯实施了一系列管控措施来应对疫情。以下两项研究关注沙特阿拉伯推文的情感分析：一项研究 COVID-19 [5], 期间关于远程学习的推文，另一项研究 COVID-19[6]期间关于政府遏制疫情的预防措施的推文。在 [6], 的研究中，作者使用朴素贝叶斯（NB）作为机器学习模型，对包含与沙特政府实施的七项公共卫生措施相关主题标签的 53,127 条推文语料库进行训练。这些措施包括：大清真寺关闭、卡提夫封锁、学校和大学停课、购物中心、公园和餐馆关闭、体育赛事暂停、集体礼拜和每周周五聚礼暂停，以及全国宵禁。分析结果表明，沙特阿拉伯的 Twitter 用户支持政府实施的各项措施，只有一项措施例外：关闭购物中心、公园和餐馆。对于其他六项措施，结果显示正向情感多于负向情感，尤其是涉及宗教活动的措施。

[2] http://www.nlpchallenge.cerist.dz

[5]开展的研究显示了沙特阿拉伯对远程学习实施措施的接受程度。他们的情感分类器采用知名的机器学习和特征提取技术构建。最佳结果是使用单元语法和词频-逆文档频率方法进行特征提取的逻辑回归（LR），准确率达到 0.89。研究得出的结论是，正面意见多于负面意见。

Manal M.A. 等人所述的研究 [7] 关注如何通过电子学习缓解 COVID-19。这些作者还使用了传统机器学习方法，例如支持向量机、K近邻（KNN）、多项式朴素贝叶斯（MNB）和逻辑回归（LR）。通过 Twitter API 从 Twitter 提取的数据经过预处理，并根据使用 NRC 词语-情绪词典（加拿大国家研究委员会）确定的八种情绪权重进行标注。结果表明，支持向量机分类器表现最佳，准确率为 0.89。

关于阿拉伯语，COVID-19 期间发表的有关假新闻检测的研究和数据集很少，与其他语言，尤其是英语相比，相关研究尚未达到同等水平[8]。在下文中，我们将介绍 COVID-19 期间一些最新的阿拉伯语假新闻自动检测系统。Ahmed .R.M 等人的研究 [8] 介绍了一个用于检测阿拉伯语推文中 COVID-19 假新闻的公开数据集。作者比较了两个构建的语料库：2500 条人工标注的推文，以及从 700 万条提取的推文中自动标注的 34529 条推文。采用传统机器学习方法后，结果显示，自动标注和人工标注的语料库 F1分数分别为 0.93 和 0.87。

在 [9], 中，作者在四个数据集——Covid19Fakes、AraNews、讽刺新闻和 ANS——上测试了采用新策略“JointBERT”的 Transformer模型。他们提出了一种检测阿拉伯语假新闻的新技术。与阿拉伯语假新闻检测基线模型相比，该技术的表现提升了近 10%。作者使用所提出的“JointBERT”方法进行序列标注和编码过程，关于该模型的更多细节见[10]。

在[11], 中，作者构建了两个与 COVID-19 相关的阿拉伯语假新闻检测数据集。第一个数据集包含 16,000 个样例，由“Kaggle”数据翻译成英语；第二个数据集包含 10,000 个样例，收集自阿拉伯语网站和推文。作者开发了八种 Transformer模型，包括 Giga-Bert、Roberta-Base、AraBert、Arabic-BERT、ARBERT、MarBert、Araelectra 和 QaribBert，以及上下文化阿拉伯语嵌入模型。与翻译而来的数据集相比，这些模型在阿拉伯语数据集上的表现要好得多。ARBERT 预训练模型的准确率达到 0.98。

3. 实验

In t本节介绍构建情感分析模型和FN检测模型所遵循的步骤 .

3.1. 预训练模型描述

Transformer是一种简单的网络架构，由 [12], 提出，基于注意力机制 [13]。注意力机制背后的主要思想是，通过扩展向量长度 [14]，使编码器-解码器模型能够自动从输入源中搜索相关部分，以便为机器翻译预测相关的目标词。

BERT [15]是一种语言表示。对于文本分类，BERT-base 模型包含一个由 12 个 Transformer 块组成的编码器。BERT 接受一个长度不超过 512 的

个词元序列作为输入，并输出该序列的表示 [16]。预训练的 BERT 模型在多项自然语言处理任务中达到最先进水平，例如问答和语言推理 [15]。

我们使用 Hugging Face 库[3]中的预训练模型 bert-mini-arabic 和 bert-base-arabic[17]，分别用于情感分析任务和假新闻检测任务。该模型使用约 82 亿个单词进行预训练。该模型是卷积神经网络（CNN）与 BERT [17]的结合。

3.2. 数据预处理

我们在预处理数据之前先分析数据分布。如以下表格所示（见表1），情感分析任务和假新闻检测任务的训练数据集类别分布均衡，这可能会影响模型的准确率表现。

我们采用了一些常用设置来准备用于 BERT 模型的训练数据集。鉴于 BERT 模型高度依赖句子中词语的上下文，因此在数据清洗过程中，我们尽量保留每个句子（“推文”）的语义和句法结构。

我们使用 Python 的“String”模块和正则表达式移除网址及标点符号。我们将文本中的某些标点符号替换为空格，并删除表情符号。数字则保持不变。请注意，我们没有发现重复推文，并且对两个数据集都执行了上述预处理。最后，数据已准备就绪，可用于使用 BERT 预训练模型进行训练。

表1数据集分布

Task	各类别的计数值
情感分析（情感分析）	中性情感 1600 正向情感 1571 负向情感 957
总计	4128 条推文
虚假新闻（FN）	无法判断 5024 也许 1471 是 367 否 1799
总计	8661条推文

如表1所示，情感分析任务中推文总语料库由38%的中性和积极情感组成。在虚假新闻任务中，“无法判断”类别占58%，而“是”和“否”类别分别占语料库的4%和16%。由于不平衡数据集会影响大多数机器学习算法，因此预测少数类可能较为困难 [18]。

[3]https://huggingface.co/

3.3. 实验性微调

微调过程是使用BERT等预训练模型时的关键步骤。该过程包括调整预训练模型的参数，使其适用于所应用的特定任务。对于BERT模型，在我们的案例中，我们使用了以下微调参数：

- 训练批量大小：训练过程中，我们使用了 10 的批量大小。批量大小是指在更新模型权重之前一次处理的样本数量。
- 学习率：学习率决定优化器更新模型权重时所采用的步长。对于我们的模型，我们使用了 2e-5 的学习率。
- 优化器：训练过程中，我们使用 AdamW 优化器来更新模型权重。AdamW 是Adam优化器的一种变体，非常适合对 BERT 等预训练模型进行微调。
- 训练轮次：训练轮次是指模型在整个数据集上训练的次数。对于我们的模型，总计进行了 10 个训练轮次。
- 最大序列长度：最大序列长度是指一次输入 BERT 模型的词元最大数量。在我们的实验中，最大序列长度设为 256。

通过使用这些微调参数，我们能够有效地对预训练 BERT 模型进行微调，以用于情感分析和假新闻检测任务。

3.4. 结果与讨论

为了评估模型的性能，我们使用经过微调的 BERT 模型预测情感分析任务和假新闻检测任务的目标，并将结果提交给挑战赛委员会。

由于尚未获得挑战赛其他参与者的结果，我们无法将我们的结果与之比较。尽管如此，我们可以说，在情感分析和假新闻检测方面，使用预训练语言模型 BERT 能取得良好的结果。

如以下表格所示（表2 和表3），就三项性能指标而言，情感分析任务的结果优于假新闻检测任务。这一结果可归因于情感分析任务数据集中的类别平衡分布。

FN检测数据集是不平衡的，这可能导致输出产生偏差，并影响模型的整体性能。

表2Techno 团队在验证数据集上针对情感分析和假新闻检测任务 task1.c 的结果

Task	精确率	召回率	F1 分数
SA	0.94	0.93	0.94
FN	0.87	0.86	0.86

表3Techno 团队在测试数据集上针对 task1.c 的情感分析和虚假新闻检测任务的结果

Task	准确率	F1 分数
情感分析（运行1）	0.9332	0.9332
情感分析（运行2）	0.9332	0.9331
情感分析（运行3）	0.9381	0.9381
FN	0.9021	0.9016

4. 结论

在这项工作中，我们利用自然语言处理和预训练语言模型 BERT，完成了 CERIST'22 共享任务"task1.c"中的两个分类子任务。在情感分析任务中，我们获得了 0.93 的 F1 分数，在假新闻检测任务中获得了 0.90 的 F1 分数。这些结果表明，该方法有效解决了与 COVID-19 相关的情感分析任务。然而，假新闻检测任务的不平衡数据集凸显了该模型的局限性。为了进一步提升结果，在未来的工作中，我们计划采用其他学习方法。

参考文献

[1] M. A. Hassonah, R. Al-Sayyed, A. Rodan, A.-Z. Ala'M, I. Aljarah, H. Faris, 《一种用于 Twitter 上各类主题情感分析的高效混合过滤器与进化式封装器方法》, Knowledge-Based Systems 192 (2020) 105353。[2] K. Stahl, 《社交媒体中的假新闻检测》, 加利福尼亚州立大学斯坦尼斯洛斯分校 6 (2018) 4–15。[3] K. Shu, A. Sliva, S. Wang, J. Tang, H. Liu, 《社交媒体中的假新闻检测：数据挖掘视角》, ACM SIGKDD explorations newsletter 19 (2017) 22–36。[4] S. Alhumoud, 《使用深度学习对 COVID-19 Twitter 数据进行阿拉伯语情感分析》, International Journal of Co mputer Science and Network Security (2020) 132–138。[5] M. Aljabri, S. M. B. Chrouf, N. A. Alzahrani, L. Alghamdi, R. Alfehaid, R. Alqarawi, J. Alhuthayfi, N. Alduhailan, 《COVID-19 疫情期间沙特阿拉伯远程学习相关阿拉伯语推文的情感分析》, Sensors 21 (2021) 5431。[6] M. Alhajji, A. Al Khalifah, M. Aljubran, M. Alkhalifah, 《沙特阿拉伯推文中关于遏制 COVID-19 的政府预防措施的情感分析》(2020)。[7] M. M. Ali, 《关于在线学习以减轻 COVID-19 影响的阿拉伯语情感分析》, Journal of Intelligent Systems 30 (2021) 524–540。[8] A. R. Mahlous, A. Al-Laith, 《COVID-19 疫情期间阿拉伯语推文中的假新闻检测》, International Journal of Advanced Computer Science and Applications 12 (2021)。[9] W. Shishah, 《用于检测阿拉伯语假新闻的 JointBERT》, IEEE Access 10 (2022) 71951–71960。

[10] W. Shishah, 《使用联合学习的 BERT 模型进行假新闻检测》, 《阿拉伯科学与工程杂志》46 (2021) 9115–9127。[11] A. B. Nassif, A. Elnagar, O. Elgendy, Y. Afadar, 《基于深度上下文化嵌入模型的阿拉伯语假新闻检测》, 《神经计算与应用》(2022) 1–14。[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polo-sukhin, 《Attention is All You Need》, 《神经信息处理系统进展》30 (2017)。[13] D. Bahdanau, K. Cho, Y. Bengio, 《通过联合学习对齐与翻译实现神经机器翻译》, arXiv 预印本 arXiv:1409.0473 (2014)。[14] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, 《用于统计机器翻译的循环神经网络编码器-解码器短语表示学习》, arXiv 预印本 arXiv:1406.1078 (2014)。[15] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, 《BERT：用于语言理解的深度双向变换器预训练》, arXiv 预印本 arXiv:1810.04805 (2018)。[16] C. Sun, X. Qiu, Y. Xu, X. Huang, 《如何微调 BERT 以进行文本分类？》, 载于《中国计算语言学大会》, 施普林格, 2019年, 第 194–206 页。[17] A. Safaya, M. Abdullatif, D. Yuret, 《KUISAIL 在 SemEval-2020 任务 12中的系统：用于社交媒体攻击性言论识别的 BERT-CNN》, 载于《第十四届语义评估研讨会论文集》, 2020年, 第 2054–2059 页。[18] T. Zhang, F. Wu, A. Katiyar, K. Q. Weinberger, Y. Artzi, 《重新审视少样本 BERT 微调》, arXiv 预印本 arXiv:2006.05987 (2020)。