arXiv:2301.00321v1 [cs.CL] 1 Jan 2023

使用自然语言处理技术判断洪水相关性并从 Twitter 帖子中识别地点

穆罕默德 苏莱曼[1], †, 穆罕默德阿西夫[1], †, 塔亚布 扎米尔[2], †, 阿亚兹 梅赫穆德[1], †, 杰布兰汗[3], 纳西尔艾哈迈德[1] 和 卡希夫 艾哈迈德[4]

[1]工程与技术大学计算机科学与工程系，白沙瓦，巴基斯坦[2]阿巴辛大学伊斯兰堡校区，巴基斯坦[3]亚洲大学人工智能系，韩国[4]明斯特理工大学计算机科学系，科克，爱尔兰

摘要

本文介绍了我们为 MediaEval 2022 DisasterMM 任务提出的解决方案。该任务由两个子任务组成，分别是：（i）Twitter 帖子相关性分类（RCTP），以及（ii）推文文本中的位置提取（LETT）。RCTP 子任务旨在区分与洪水相关的社交媒体帖子和无关帖子，而 LETT 是一项命名实体识别（NER）任务，旨在从文本中提取位置信息。针对 RCTP，我们提出了四种基于 BERT、RoBERTa、DistilBERT 和 ALBERT 的不同方案，分别取得了 0.7934、0.7970、0.7613 和 0.7924 的 F1分数。针对 LETT，我们使用了三种模型，即 BERT、RoBERTa 和 Distil BERTA，分别取得了 0.6256、0.6744 和 0.6723 的 F1分数。

1. 引言

自然灾害是通常由地球物理、水文、气候及气象因素引发的危险事件。这些危险事件可能对人类生命和基础设施造成不利影响。洪水便是其中一种，且频繁发生于世界各地。与其他自然灾害类似，洪水可能对公共卫生和基础设施造成重大影响。例如，许多情况下都发现，道路和通信基础设施在洪水期间遭到严重破坏 [1]。

迅速有效地应对灾害有助于减轻其不利影响。获取相关且及时的信息对于有效应对至关重要。文献指出，在多种情况下，由于若干原因，获取相关信息可能受到影响，例如当地没有记者，以及通信设施受损 [2]。近年来，社交媒体和众包已被用作紧急情况下的通信、信息收集与传播渠道。为此，人们提出了多种有趣的解决方案，用于收集、分析社交媒体内容并从中提取有价值的见解

[2]。然而，社交媒体内容也存在一些局限性。例如，社交媒体内容通常比较嘈杂，因此很难获取相关信息。同样，

MediaEval'22：多媒体评估研讨会，*2023*年*1*月*13*日至*15*日，挪威卑尔根及线上 *通讯作者。

†这些作者贡献相同。 kashif.ahmad@mtu.ie (K. Ahmad)© 2022 本文作者版权所有。允许依据知识共享署名 4.0 国际许可协议 (CC BY 4.0) 使用。CEUR 研讨会论文集 (CEUR-WS.org)



对于所有相关帖子而言，作为判断内容相关性的关键信息，地理位置信息未必都能获取。

鉴于社交媒体内容在灾害分析中的重要性及其应用，多年来，洪水检测也被纳入 MediaEval 基准竞赛，作为一项共享任务。本文介绍了针对 MediaEval 中 MMDisaster 任务提出的解决方案 2022 [3]。该挑战旨在解决社交媒体灾害分析中的两项关键难题。第一个子任务旨在通过自动筛选社交媒体内容，获取相关内容，从而减少社交媒体噪声。第二个子任务旨在从社交媒体文本中提取位置信息，从而自动定位洪水引发的潜在事件。针对这两个子任务，我们提出了若干有趣的解决方案，详见第 3 节。

2. 相关工作

近年来，社交媒体在不同应用领域的潜力已得到广泛探索 [4, 5]。社交媒体内容已被证明非常有效的一些关键应用包括公共卫生 [6], 、教育 [7], 和公共资源管理 [8]。社交媒体平台也已被广泛应用于灾害和紧急情况中的各类任务 [2]。例如，Hao 等人 [9] 提出了一种多模态框架，利用多种社交媒体图像和文本信息，对受灾地区进行损害评估。该研究分析的关键因素包括灾害类型、严重程度和损害类型。Wu 等人 [10] 也利用社交媒体数据及其相关地理位置信息，在灾害发生后生成预警并开展损害评估分析。另一方面，Ahmad 等人 [11], 利用社交媒体图像分析洪水后的道路状况。具体而言，作者提出了一种早期与晚期融合框架，用于识别洪水淹没地区的可通行道路。Alam 等人 [12] 探索了社交媒体内容在另一项相关任务中的应用，即评估洪水严重程度。为此，作者收集了一个名为 CrisisMD 的大规模基准数据集。该数据集汇集了大量包含文本和视觉内容的 Twitter 帖子。Hassan 等人 [13] 则通过从社交媒体平台分享的视觉内容中提取情感和情绪，探索了自然灾害中一个略有不同的方面。作者详细阐述了如何利用对灾害相关社交媒体视觉内容进行的视觉情感分析，为新闻机构、公共机构和人道主义组织等不同利益相关者提供帮助。

尽管社交媒体内容已被证明在灾害分析的不同任务中非常有效，但它仍存在一些局限性，例如噪声数据以及无法获取地理位置信息。本文提出了一种解决方案，以克服这些挑战。

3. 方法

3.1. Twitter 帖子相关性分类（RCTP）

首先，我们分析了现有的多媒体内容。在分析过程中，我们发现大多数帖子缺少视觉内容。此外，大多数图像都无关。因此，我们决定在解决方案中仅使用文本信息。我们的 RCTP 子任务框架由两个步骤组成。第一步，我们采用了一些预处理技术，通过删除用户名、网址、表情符号和停用词等不必要的信息来清理数据。

预处理后，我们使用了多种先进的自然语言处理算法，包括 BERT [14], RoBERTa [15], DistilBERT [16], 和 ALBERT [17] ，对文本进行分类。由于这是一项二元分类任务，所有方法的损失函数都基于二元交叉熵。此外，我们使用 Adam优化器，批量大小设为 32，训练 20 个轮次。

3.2. 从 Twitter 帖子中提取地点（LETT）

LETT 子任务被视为命名实体识别（NER）任务。命名实体识别旨在定位文本中的命名实体，并将其归入预先定义的类别 [18]。在此任务中，我们关注识别表示文本序列中地点起始词及后续词的词语。在 LETT 中，标注以词为单位。与 RCTP 任务类似，我们采用了多种最先进的算法，包括 BERT、RoBERTa、DistilBERT 和 ALBERT。值得注意的是，在此任务中，由于标注以词为单位，因此在训练模型之前，我们没有使用任何预处理技术。

3.3. 数据集

两个子任务均发布了独立的数据集。RCTP 子任务的数据集共包含 8,000 条推文。这些推文于 2020年5月25日 至 2020年6月12日 期间收集，使用的关键词为意大利语中的洪水相关词汇，如“alluvione”、“allagamento”和“esondazione”。该数据集分为开发集和测试集。开发集包含 5,337 条推文，测试集共包含 1,315 条推文。

LETT子任务的数据集由约 6,000 条推文组成，这些推文使用与洪水相关的意大利语关键词，于2017年3月25日至2018年8月1日期间收集。此子任务的标注以推文中的词语为单位提供。

4. 结果与分析

4.1. RCTP子任务的运行说明

表1 展示了所提出方案在开发集上的实验结果。需要指出的是，在开发集上的实验中，我们使用开发集的 70% 样本进行训练、20% 进行测试，10% 样本用于验证。如表中所示，模型在干净数据集和不干净数据集上的表现没有显著差异。就各个模型的表现而言，与其他模型相比，BERT 的结果略好。表2 给出了所提出方案在测试集上的官方结果。需要指出的是，在测试集实验中，模型使用完整的开发集进行训练。本任务总共提交了 4 组不同的运行结果。第一、第二和第四组分别基于使用不干净数据集训练的 BERT、RoBERTa 和 DistilBERT 模型。第三组则基于使用干净数据集训练的 BERT 模型。使用不干净数据集训练的模型表现优于使用干净数据集训练的模型。这表明，预处理信息导致一些相关特征被移除，从而对结果产生了负面影响。

表 1RCTP任务在开发集上的实验结果。

模型	干净数据集上的F1分数	不干净数据集上的F1分数
BERT	0.95	0.94
RoBERTa	0.94	0.93
DistilBERT	0.93	0.93
ALBERT	0.92	0.92

表2RCTP任务在测试集上的实验结果。

Run	精确率	召回率	F1分数
1（BERT 在不干净数据集上的表现）	0.6949	0.9251	0.7934
2（RoBERTa 在未清洗数据集上）	0.6947	0.9347	0.7970
3（BERT 在干净数据集上）	0.6486	0.9213	0.7613
4（DistilBERT 在不干净数据集上）	0.6940	0.9232	0.7924

表 3LETT 任务在开发集上的实验结果。

模型	F1分数
BERT	0.7752
RoBERTa	0.8014
DistilBERT	0.7658
ALBERT	0.6827

表 4LETT 任务在测试集上的实验结果。

Run	精确结果			部分结果		
	精确率	召回率	F1分数	精确率	召回率	F1分数
1（BERT）	0.596	0.522	0.556	0.628	0.622	0.625
2（RoBERTa）	0.540	0.676	0.600	0.577	0.810	0.674
3（DistilBERT）	0.563	0.604	0.583	0.610	0.760	0.677

4.2. LETT 子任务的运行说明

表3 展示了所提出方案在 LETT 子任务开发集上的实验结果。与 RCTP 类似，我们使用开发集的 70% 样本进行训练、20% 进行测试，10% 样本用于验证。模型在开发集上的结果存在显著差异。总体而言，RoBERTa 的结果更好，比 BERT 模型取得的第二高结果高出 3%。表4 给出了 LETT 子任务的官方结果。在测试集实验中，模型使用完整的开发集进行训练。需要指出的是，在部分结果中，被省略的样本计为错误；而在部分结果中，被省略的样本则完全忽略，不作任何惩罚。如表中所示，总体而言，与 BERT 模型的原始实现相比，RoBERTa 和 DistilBERT 的结果更好。

5. 结论

本文介绍了我们针对 MediaEval 2022 中发布的 DisasterMM 挑战所提出的解决方案。对于两个子任务，我们采用了多种先进的自然语言处理算法。在当前实现中，所有模型均单独使用；不过，我们认为，如果通过基于优势的融合方法联合使用，这些模型可以相互补充。未来，我们计划采用不同的基于优势的融合方法，在两个子任务中联合发挥各个模型的能力。

参考文献

[1] D. T. Nguyen, F. Ofli, M. Imran, P. Mitra，在灾害期间基于社交媒体图像数据进行损害评估，载于：2017年社交网络分析与挖掘进展国际会议论文集，2017年，第569–576页。[2] N. Said, K. Ahmad, M. Riegler, K. Pogorelov, L. Hassan, N. Ahmad, N. Conci，社交媒体和卫星图像中的自然灾害检测：综述，Multimedia Tools and Applications 78 (2019) 31267–31302。[3]S. Andreadis, A. Bozas, I. Giala mpoukidis, A. Moumtzidou, R. Fiorin, F. Lombardo, T. Mavropoulos, D. Norbiato, S. Vrochidis, M. Ferri, I. Kompatsiaris，DisasterMM：MediaEval 2022灾害相关社交媒体数据任务中的多媒体分析，载于：MediaEval 2022研讨会论文集，挪威卑尔根及线上，2023年。[4] K. Ahmad, K. Pogorelov, M. Riegler, N. Conci, P. Halvorsen，社交媒体与卫星，Multimedia Tools and Applications 78 (2019) 2837–2875。[5] I. Alsmadi, K. Ahmad, M. Nazzal, F. Alam, A. Al-Fuqaha, A. Khreishah, A. Algosaibi, 社交网络应用的对抗性自然语言处理：攻击、防御与研究方向，IEEE计算社会系统汇刊 (2022)。[6] S. Andreadis, G. Antzoulatos, T. Mavropoulos, P. Giannakeris, G. Tzionis, N. Pantelidis, K. Ioannidis, A. Karakostas, I. Gialampoukidis, S. Vrochidis等, 通过利用自动添加地理标签的推文可视化COVID-19在意大利传播的社交媒体分析平台，Online Social Networks and Media 23 (2021) 100134。[7] A.-R. 等，信息系统成功和技术接受模型对教育领域社交媒体因素的影响，Sustainability 13 (2021) 7770。[8] K. Ahmad, M. Ayub, J. Khan, N. Ahmad, A. Al-Fuqaha，社交媒体作为水质反馈的即时来源，IEEE技术与社会汇刊 (2022)。[9] H. Hao, Y. Wang，利用多模态社交媒体数据快速评估灾害损害，International Journal of Disaster Risk Reduction 51 (2020) 101760。[10] D. Wu, Y. Cui，利用社交媒体数据和地理位置信息进行灾害预警与损害评估分析，Decision support systems 111 (2018) 48–59。[11] K. Ahmad, K. Pogorelov, M. Riegler, O. Ostroukhova, P. Halvorsen, N. Conci, R. Dahyot，自动检测遥感数据和社交媒体数据中洪水过后的可通行道路，Signal Processing: Image Communication 74 (2019) 110–118。[12] F. Alam, F. Ofli, M. Imran，CrisisMMD：来自自然灾害的多模态Twitter数据集，载于：第十二届AAAI网络与社交媒体国际会议，2018年。[13] S. Z. Hassan, K. Ahmad, S. Hicks, P. Halvorsen, A. Al-Fuqaha, N. Conci, M. Riegler，社交媒体灾害图像的视觉情感分析，Sensors 22 (2022) 3628。

[14] 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)。[15] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov，RoBERTa：经过稳健优化的BERT预训练方法，arXiv预印本 arXiv:1907.11692 (2019)。[16] V. Sanh, L. Debut, J. Chaumond, T. Wolf，DistilBERT：BERT的精简版本，更小、更快、更经济、更轻量，arXiv预印本 arXiv:1910.01108 (2019)。[17] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, R. Soricut，ALBERT：用于语言表征自监督学习的轻量级BERT，arXiv预印本 arXiv:1909.11942 (2019)。[18] J. Li, A. Sun, J. Han, C. Li，命名实体识别深度学习综述，IEEE知识与数据工程汇刊 34 (2020) 50–70。