arXiv:2301.00418v1 [cs.CL] 1 Jan 2023

越南语情感分类是否需要分词?

Duc-Vu Nguyen[1, 2], Ngan Luu-Thuy Nguyen[1, 2][1]越南胡志明市信息技术大学[2]越南胡志明市越南国家大学
{vund,ngannlt}@uit.edu.vn

摘要——据我们所知，本文首次尝试回答越南语情感分类是否需要分词。为此，我们提出了五种越南语预训练单语 S4 语言模型，其中一种不进行分词，另外四种则在数据预处理阶段使用 RDRseg-menter、 uitnlp、 pyvi或 underthesea工具包。根据在两个语料库上进行的全面实验结果——包括来自新闻和社交媒体的技术文章评论 VLSP2016-SA 语料库，以及教育调查 UIT-VSFC 语料库——我们提出两点建议。首先，对于朴素贝叶斯或支持向量机等传统分类器，来自社交领域的越南语情感分类语料库可能不需要分词。其次，在使用 BPE 方法并输入深度学习模型之前，越南语情感分类需要进行分词。如此一来，RDRsegmenter便是 uitnlp、 pyvi和 underthesea等工具包中稳定的分词工具包。

索引词—自然语言处理、越南语情感分类、越南语分词、结构化状态空间序列（S4）

I. 引言

分词是越南语自然语言处理领域的一项基础问题。例如，越南语文本“hiện đại hóa đất nước”（使国家现代化hiện_đại_hoá 国家đất nước），由五个音节组成，可切分为“hiện_đại_hoáđất_nước”。下划线表示空格作为音节分隔符，而空格也用于分隔词。这是越南语自然语言处理早期面临的一项挑战[1]。越南语句法分析中的常见任务，如词性标注 [2], 成分句法分析 [3, 4], 依存句法分析 [5, 6, 7], 和语义分析[8], , 都需要先进行分词。在这些任务中，分词错误会直接影响任务表现。因此，许多以往研究提出了各种方法来提升越南语分词性能，包括仅针对分词的单任务方法 [1, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18]以及包含分词和词性标注的多任务方法 [19, 20], 依存句法分析 [21]。从上述以往研究可以看出，分词是越南语句法分析任务（包括词性标注、成分句法分析、依存句法分析和语义分析）不可或缺的步骤。

除了越南语句法分析任务外，情感分类也吸引了许多研究人员，因为它能够迅速应用于现实生活中的应用场景。首先，Kieu 和 Pham [22] 引入了一个系统，可将计算机产品评论分类为正面或负面。Kieu 和 Pham [22] 的研究是首个在越南语中进行句子级别情感分析的工作。另一个例子是 Duyen 等人 [23], 的研究，他们提出了一个框架，可将酒店评论分为正面、中性或负面。几乎所有研究都在探讨解决情感分类任务的方法之前，先在数据预处理中进行分词。Vo 等人 [25], 的工作使用了 vnTokenizer工具包 [13] ，对 VLSP2016-SA [24]和 VS语料库进行预处理; Duyen 等人 [23],提出的酒店评论语料库；Bach 等人[26], 提出的电子设备语料库；Vo 等人[27], 提出的教育调查语料库；以及 Tran 和 Phan [28]提出的在线评论语料库。Pham 等人 [19], 提出的另一种越南语分词器，被 Ha 等人 [29], 、Kieu 和 Pham [22], 以及 Vu 等人[30] 用于产品评论领域的研究。之后，Phan 和 Cao [31] 使用 JVnSegmenter工具包 [11] 对他们的食品评论语料库进行预处理。此外，Pham 等人 [32] 使用 UETsegmenter [15] 对他们的移动产品评论语料库进行预处理。最后，关于学生反馈情感分析的研究[33, 34] [35] 以及关于情绪识别的研究 [36] ，都使用了 Nguyen 等人 [16] 的 RDRsegmenter开展研究。上述所有越南语情感分析研究均使用了发表于科学论文中的越南语分词器。此外，我们还可以考察两种著名的越南语分词器，即 pyvi[1] 和 underthesea[2] ；截至目前，它们尚未在科学论文中发表。例如，Van Thin 等人 [37] 关于 VLSP2018-SA 语料库 [24] 情感分析的研究，以及 Nguyen 等人 [38] 关于产品评论的研究，都使用了 pyvi 工具包。另一个例子是，Nguyen 等人 [39] 使用 underthesea 工具包对其电子产品评论数据集进行预处理。最后，我们发现，许多先前的越南语情感分类研究都在预处理阶段进行了分词。

[1]https://pypi.org/project/pyvi/[2]
https://pypi.org/project/underthesea/

尽管我们可以看到，分词阶段对于越南语情感分类问题几乎必不可少，但近期已有一些研究在不进行分词的情况下探讨越南语情感分类问题。例如，相关研究[40] 使用 fastText [41] 模型对输入数据进行预处理和嵌入，且未进行越南语分词。另一些例子中，相关研究 [42, 43, 44] 使用预训练多语言语言模型 mBERT 研究情感分类问题 [45]，，该模型不需要进行越南语分词。据我们所知，研究 [40, 42, 43, 44] 采用子词 [46]方法，以更好地处理未见词。另一方面，fastText 和 mBERT 模型面向多语言用途。因此，研究 [40, 42,43, 44] 在越南语情感分类问题中没有使用分词阶段。事实上，研究 [34] 使用了 RDRsegmenter 工具包进行数据预处理，然后使用预训练的单语言 PhoBERT 模型 [47]，；该模型专为越南语设计，并采用字节对编码（BPE）方法 [48] 生成越南语的子词表示。

从上述观察可以看出，我们不得不承认，分词对于越南语句法分析任务至关重要，包括词性标注、成分句法分析、依存句法分析和语义分析。这也充分说明了以往研究提出众多越南语分词方法的动因。在越南语树库语料上进行分词时， [3], 跨度标注方法使用 XLM-RoBERTa 达到了最高的 F 分数 98.31% [18] [49]，，但在中央处理器设备上进行推理时速度很慢。然而，知名的越南语分词器工具包 RDRsegmenter采用规则方法，达到了 97.90% 的 F 分数，且推理速度很快。此外，从上述观察还可以看出，以往许多关于越南语情感分类的研究都使用了分词，但也有一些研究没有使用分词。因此，我们提出了一个研究问题：**“越南语情感分类是否需要分词？”**

为尝试回答上述问题，我们使用了四种越南语分词工具包[3]。首先，我们选择了快速且准确的越南语分词工具包 RDRsegmenter [16]。其次，我们选择了越南语分词工具包 uitnlp [17]，该工具包旨在消解歧义并捕获后缀。最后，我们选择了两种知名的越南语分词工具包，包括 pyvi 和 un-derthesea，它们迄今尚未经过科学发表。关于语料库，我们选择了两个与越南语情感分类问题相关的语料库，其中包括 VLSP2016-SA 语料库 [24]，该语料库包含来自新闻和社交媒体的技术文章评论，以及 UIT-VSFC 语料库，该语料库收录了教育调查[35]。关于分类器，我们选择了两种传统分类器作为基线，包括朴素贝叶斯（NB）和支持向量机（SVM），以及近期较新的文本编码器，即结构化状态空间序列模型（S4）[50]。

综上所述，我们的贡献如下：

- 我们为越南语提供了五种预训练单语 S4 语言模型，其中一种模型不进行分词，另外四种模型在预处理数据阶段使用RDRsegmenter [16]、uitnlp [17]、 pyvi或underthesea 工具包。
- 根据在两个语料库上进行的大量实验结果，包括来自新闻和社交媒体的技术文章评论所构成的 VLSP2016-SA 语料库，以及教育调查所构成的 UIT-VSFC 语料库，我们提出两点建议：
  - 对于来自社交领域的越南语情感分类语料库，使用朴素贝叶斯或支持向量机等传统分类器时，分词可能并非必要。
  - 当将分词作为使用 BPE 方法并输入深度学习模型之前的预处理步骤时，越南语情感分类需要进行分词。因此，在 uitnlp、 pyvi和 under-thesea 工具包中， RDRseg-menter 是稳定的分词工具包。

II. 用于越南语的预训练单语 S4 语言模型

本文着重探究预处理数据阶段采用不同分词工具包，对深度学习模型的越南语情感分类性能所产生的影响。因此，本节介绍了采用不同分词工具包、使用相同 S4 语言模型架构，并以相同的预训练数据和优化算法进行训练的模型。

A. 架构

今年，Gu 等人 [50] 指出，结构化状态空间序列模型（S4）有望成为一种有效的通用序列建模解决方案。因此，我们选择了 S4 模型 [50] 来开发面向越南语的预训练单语 S4 语言模型。此外，我们需要训练多达五种面向越南语的单语 S4 语言模型，包括一种不使用分词的模型，以及四种在数据预处理阶段使用RDRsegmenter、 uitnlp、 pyvi或 underthesea工具包的模型。因此，我们仅使用两个 S4 层块（而非原研究 [50]中的 16 个块），并与逐位置前馈层交替排列，特征维度为 256。遵循 [50]，，我们在 S4 线性层后使用 GLU 激活函数，并在每个块中使用两个 S4 层。嵌入

[3]由于时间有限，我们只选取了最近发布的越南语分词工具包。未来的工作中，我们将扩展研究范围，考察其他工具包。

和 softmax 层采用了来自[51] 的自适应嵌入，并设置了 5000、10000、15000 的自定义截断值。嵌入维度为 256，词汇表大小为 64K，详见下一小节（II-B）。因此，每个此类模型约有 2000 万个参数，比 PhoBERT$_{base}$[47]小约六倍。

B. 预训练数据

我们使用了一个 10 GB 的预训练数据集，该数据集由两个语料库合并而成：第一个是越南语维基百科语料库[4] (～ 1 GB)，第二个语料库(～ 9 GB) 来自 18.6 GB 的越南语新闻语料库[5]。为了公平比较上一小节（II-A）提到的五个越南语预训练单语 S4 语言模型，在完成分词或不分词的预处理后，我们使用 BPE 方法 [48] 将预训练数据集中的句子切分为子词单元，词汇表包含 64K 种子词类型。

C. 优化

我们沿用了 Gu 等人 [50] 研究中的 S4 语言模型实现[6] 。训练和评估五个越南语预训练单语 S4 语言模型时，我们将最大长度设为 512 个子词 token。按照 Gu 等人[50]，的方法，我们使用 AdamW [52] 优化模型，采用单个余弦学习率周期，最多训练 40 轮，预热步数为 10000。初始学习率设为 0.0005。在 V100 GPU（16 GB）上，我们使用批次大小 128，并将梯度累积步数设为 16。训练在 Google Colaboratory[7] 上进行。最后，除预处理阶段未进行分词的模型训练了 14 天外，其余每个 S4 语言模型均训练了 10 天。

III. 实验设置

我们评估了两种传统分类器（朴素贝叶斯和支持向量机）以及一种深度学习分类器（基于 S4 的语言建模）在两个知名越南语情感语料库上的表现，并比较是否经过分词阶段的影响。这两个语料库分别是包含来自新闻和社交媒体的技术文章评论的 VLSP2016-SA 语料库，以及教育调查的 UIT-VSFC 语料库。本文再次说明，我们关注的是在数据预处理阶段应用或不应用分词阶段对深度学习模型的越南语情感分类表现所产生的影响。这意味着我们并不着眼于取得新的最先进结果。

A. 语料库

表 I VLSP2016-SA 语料库中各类别的样本数量

	负面	中性	正面	总计
训练	1,526	1,519	1,530	4,575
验证	174	181	170	525
Test	350	350	350	1,050

表 I 展示了 VLSP2016-SA[24] 语料库的统计数据。原始语料库没有验证集。因此，我们将原始训练集划分为新的训练集和验证集。

表 II UIT-VSFC 语料库中各类别的样本数量

	负面	中性	正面	总计
训练	5,325	458	5,643	11,426
验证	705	73	805	1,583
Test	1,409	167	1,590	3,166

表II 展示了 UIT-VSFC [35]语料库的统计数据。我们按照 Nguyen 等人 [35]的划分方式划分训练集、验证集和测试集。

B. 训练

对于传统分类器，我们将一元组、二元组和三元组组合起来，作为分类问题的特征。我们使用 `scikit-learn` library [53] 来实现[8] 朴素贝叶斯和支持向量机算法。对于每个语料库，我们分别采用不进行分词以及使用 `RDRsegmenter`、 `uitnlp`、`pyvi` 或 `underthesea` 工具包进行分词的方式。使用不同的随机状态时，支持向量机的实验结果会有所不同。因此，为了公平比较，我们报告了使用 100 个随机状态进行 100 次运行的平均结果。

对于现代分类器，我们针对越南语微调了五个预训练单语 S4 语言模型，分别采用不分词以及使用适当分词工具包（`RDRsegmenter`、 `uitnlp`、 `pyvi` 或 `underthesea`）进行分词的方式。我们使用 S4 语言模型最后一层中输入句子所有位置的上下文表示的最大池化向量，作为分类的特征向量。我们使用 AdamW [52] 优化模型，并采用单个余弦学习率周期；对于 UIT-VSFC 语料库，最多训练 5 轮，对于 VLSP2016-SA 语料库，最多训练 10 轮。预热比例为 0.025。微调 UIT-VSFC 和 VLSP2016-SA 语料库时，我们将输入文本的最大长度设为 8192 个子词，批次大小设为 32。初始学习率为 0.001，AdamW 的权重衰减为 0.01。值得注意的是，由于我们在实验中没有调整超参数，因此报告 100 次运行的平均结果

采用 100 个随机状态进行公平比较。Fi 最后，我们已进行了为期 4 天的微调实验。

[4]https://github.com/NTT123 19 s18/viwik/tree/viwik https://github.com/binhvq/news-corpus#full-txttitle--description--body-v [5]https://github.com/binhvq/news-corpus#full-txttitle--description--body-v1 [6]https://github.com/NTT123/lm_with_gated_cnn_and_gpt2 [7]https://colab.research.google.com/

[8] https://scikit-learn.org/stable/

IV. 实验结果

A. 主要结果

表 III 测试集 VLSP2016-SA 语料库上分类模型 100 次运行的平均评估（%）

模型	Word 分词器	VLSP2016-SA		
		微观-F_1	宏观-F_1	加权-F_1
NB	None	**67.24**	**67.38**	**67.38**
	RDRsegmenter	66.38	66.39	66.39
	uitnlp	66.00	66.00	66.00
	pyvi	67.14	67.13	67.13
	underthesea	65.62	65.62	65.62
SVMs	None	**67.96**	**67.99**	**67.99**
	RDRsegmenter	67.56	67.56	67.56
	uitnlp	67.33	67.33	67.33
	pyvi	67.93	67.88	67.88
	underthesea	66.77	66.78	66.78
S4	None	67.44	67.42	67.42
	RDRsegmenter	68.14	68.07	68.07
	uitnlp	**68.51**	**68.47**	**68.47**
	pyvi	67.88	67.83	67.83
	underthesea	67.85	67.85	67.85

表 III 展示了在 VLSP2016-SA 语料库上的平均评估结果。首先，未经过分词阶段训练的朴素贝叶斯分类器，其表现优于经过分词阶段训练的其他朴素贝叶斯分类器。支持向量机分类器也出现了同样的情况。其次，经过分词阶段训练的基于 S4 的分类器，其表现优于未经过分词阶段训练的基于 S4 的分类器。特别是，使用 `uitnlp` 工具包的基于 S4 的分类器，其结果优于 `RDRsegmenter`、`pyvi`和 `underthesea`。

表 IV UIT-VSFC 语料库测试集上分类模型 100次运行的平均评估（%）

模型	Word 分词器	UIT-VSFC		
		微观-F_1	宏观-F_1	加权-F_1
NB	None	86.26	59.83	84.07
	RDRsegmenter	86.99	60.32	84.77
	uitnlp	87.11	60.79	84.95
	pyvi	**87.18**	**60.83**	**84.99**
	underthesea	86.99	60.33	84.78
SVMs	None	89.17	73.48	88.55
	RDRsegmenter	**89.36**	72.96	88.60
	uitnlp	89.35	73.05	**88.62**
	pyvi	89.29	**73.10**	88.54
	underthesea	89.27	72.58	88.47
S4	None	90.88	75.82	90.11
	RDRsegmenter	**91.62**	**77.84**	**90.99**
	uitnlp	91.36	77.17	90.68
	pyvi	91.40	77.74	90.80
	underthesea	91.35	77.26	90.72

表 IV 展示了 UIT-VSFC 语料库上的平均评估结果。正如表 IV所示，使用分词阶段的朴素贝叶斯、支持向量机和基于 S4 的分类器，其性能高于未使用分词阶段的其他分类器。特别是，使用 `RDRsegmenter` 工具包的基于 S4 的分类器，其结果高于 `uitnlp`、`pyvi`和 `underthesea`。

B. 讨论

正如表III所示，在越南语情感分类任务中，对朴素贝叶斯和支持向量机等传统分类器使用分词，其性能可能低于不使用分词的分类器。根据 Nguyen 等人的研究 [17], , `uitnlp` 在越南语树库语料上进行分词时，其 F 分数高于 `RDRseg-menter` [16] 的分词结果 [3]。然而，VLSP2016-SA 语料库包含来自新闻和社交媒体的技术文章评论。因此，社交语料库的分词标准可能不同于越南语树库语料的分词标准 [3]。最后，传统分类器使用 `uitnlp` 工具包时性能低于使用 `RDR`的分类器，可能是因为 `uitnlp` 工具包在越南语树库语料上进行分词训练时出现了过拟合 [3], , 当 `uitnlp` 工具包面对社交语料库时，这种情况并不理想。这表明，为社交媒体文本构建越南语分词和词性标注系统，如 Bach 等人的研究 [54] 所示，值得重视和关注。

从这种分析方式来看，在表 IV中，经过分词阶段的朴素贝叶斯、支持向量机和基于 S4 的分类器，性能优于未经过分词阶段的其他分类器。这可以解释为，教育调查领域的分词标准与越南语树库语料库的分词标准相近[3]。由于 UIT-VSFC 语料库经过人工规范化处理 [35]。

五、结论

本文试图回答分词对于越南语情感分类是否必要。根据对两个语料库进行的综合实验结果，包括来自新闻和社交媒体的技术文章评论语料库 VLSP2016-SA 语料库，以及教育调查语料库 UIT-VSFC 语料库，我们提出两点建议。首先，对于使用朴素贝叶斯或支持向量机等传统分类器的情况而言，如果越南语情感分类语料库来自社交领域，则分词可能并非必要。其次，当在使用 BPE 方法并输入深度学习模型之前进行分词时，越南语情感分类需要分词。通过这种方式，在 `uitnlp`、`pyvi`和 `underthesea`工具包中，`RDRsegmenter`是用于分词的稳定工具包，因为在我们的实验中，`RDRsegmenter`在情感分类任务上表现稳定，并且具有较快的推理速度[16]。

致谢

本研究获得了越南国家大学胡志明市信息技术大学科学研究支持基金的资助。

参考文献

[1]D. Dinh、K. Hoang 和 V.-T. Nguyen，《越南语分词》，载于第六届自然语言处理太平洋地区研讨会论文集，2001 年，第 749–756 页。[2] H.-P. Le、A. Roussanaly、T.-M.-H. Nguyen 和 M. Rossignol，《用于越南语文本词性标注的最大熵方法实证研究》，载于自然语言自动处理 - *TALN 2010*。加拿大蒙特利尔：ATALA（语言自动处理协会），2010 年，第 12 页。[3] P.-T. Nguyen、X.-L. Vu、T.-M.-H. Nguyen、V.-H. Nguyen 和 H.-P. Le，《构建大型越南语句法标注语料库》，载于第三届语言学标注研讨会（*LAW III*）论文集，2009 年，第 182–185 页。[4] Q. T. Nguyen、Y. Miyao、H. T. Le 和 N. T. Nguyen，《确保越南语树库的标注一致性与准确性》，语言资源与评测，第 52 卷，315，2018 年第 1 期，第 269 页。[5] D. Q. Nguyen、D. Q. Nguyen、S. B. Pham、P.-T. Nguyen 和 M. Le Nguyen，《从树库转换到越南语自动依存句法分析》，载于自然语言处理与信息系统，E. Métais、M. Roche 和 M. Teisseire 编。Cham：Springer International Publishing，2014 年，第 196–207 页。[6] L. N. Thi、L. H. My、H. Nguyen Viet、H. Nguyen Thi Minh 和 P. Le Hong，《构建用于越南语依存句法分析的树库》，载于*2013* 年 *RIVF* 国际计算与通信技术会议——研究、创新与未来愿景（*RIVF*），2013 年，第 147–151 页。[7] K.-H. Nguyen，《BKTreebank：构建越南语依存树库》，载于第十一届语言资源与评测国际会议论文集（*LREC 2018*）。日本宫崎：欧洲语言资源协会（ELRA），2018 年 5 月。[8] L. H. Phuong、P. Hoang、P. Khoai、N. Huyen、N. Luong 和 N. Hiep，《越南语语义角色标注》，越南国家大学科学学报：计算机科学与通信工程，第 33 卷，第 2 期，2017 年。[9] D. Dinh 和 T. Vu，《越南语分词的最大熵方法》，载于*2006* 年研究、创新与未来愿景国际会议论文集，2006 年，第 248–253 页。[10] L. A. Ha，《越南语分词方法》，载于*2003* 年语料库语言学会议论文集，2003 年，第 282–287 页。[11] C.-T. Nguyen、T.-K. Nguyen、X.-H. Phan、L.-M. Nguyen 和 Q.- T. Ha，《基于条件随机场和支持向量机的越南语分词：一项研究》，载于太平洋亚洲计算语言学会议论文集。清华大学出版社，2006 年，第 215–222 页。[12] Q.-T. Dinh、H.-P. Le、T.-M.-H. Nguyen、C.-T. Nguyen、M. Rossig- nol 和 X. L. Vu，《越南语文本分词：方法比较》，载于第十一届语言资源与评测国际会议论文集（*LREC 2008*），2008 年，第 1933–1936 页。[13] H.-P. Le、T.-M.-H. Nguyen、A. Roussanaly 和 T.-V. Ho，《越南语文本分词的混合方法》，载于语言与自动机理论及应用。施普林格柏林海德堡，2008 年，第 240–249 页。[14] Song Nguyen Duc Cong、Quoc Hung Ngo 和 R. Jiamthap- thaksin，《最先进的越南语分词》，载于*2016* 年第二届信息技术科学国际会议（*ICSITech*），2016 年，第 119–124 页。[15] T. P. Nguyen 和 A. C. Le，《越南语分词的混合方法》，载于*IEEE-RIVF* 会议论文集，2016 年，第 114–119 页。[16] D. Q. Nguyen、D. Q. Nguyen、T. Vu、M. Dras 和 M. Johnson，《快速准确的越南语分词器》，载于语言资源与评测会议论文集，2018 年，第 2582–2587 页。[17] D.-V. Nguyen、D. Van Thin、K. Van Nguyen 和 N. L.-T. Nguyen，《基于支持向量机的越南语分词：歧义消解与后缀捕获》，载于计算语言学，L.- M. Nguyen、X.-H. Phan、K. Hasida 和 S. Tojo 编。新加坡：Springer Singapore，2020 年，第 400–413 页。[18] D.-V. Nguyen、L.-B. Vo、D. V. Thin 和 N. L.-T. Nguyen，《越南语和中文分词的跨度标注方法》，载于 *PRICAI 2021*：人工智能趋势，D. N. Pham、T. Theeramunkong、G. Governatori 和 F. Liu 编。Cham: Springer International Publishing, 2021, 244–258.

[19] D. D. Pham、G. B. Tran 和 S. B. Pham，《利用词性标签进行越南语分词的混合方法》，载于*2009*年知识与系统工程国际会议，2009年，第154–161页。[20] D. Q. Nguyen、T. Vu、D. Q. Nguyen、M. Dras 和 M. Johnson，《从越南语分词到词性标注》，载于*2017*年澳大拉西亚语言技术协会研讨会论文集，2017年，第108–113页。[21] D. Q. Nguyen，《用于越南语分词、词性标注和依存句法分析的神经联合模型》，载于澳大拉西亚语言技术协会第*17*届年度研讨会论文集，2019年，第28–34页。[22] B. T. Kieu 和 S. B. Pham，《越南语情感分析》，载于*2010*年第二届知识与系统工程国际会议，2010年，第152–157页。[23] N. T. Duyen、N. X. Bach 和 T. M. Phuong，《越南语情感分析的实证研究》，载于*2014*年先进通信技术国际会议（*ATC 2014*），2014年，第309–314页。[24] H. T. M. Nguyen、H. V. Nguyen、Q. T. Ngo、L. X. Vu、V. M. Tran、B. X. Ngo 和 C. A. Le，《VLSP 共享任务：情感分析》，《计算机科学与控制论期刊》，第34卷，第4期，第295–310页，2019年1月。[25] Q.-H. Vo、H.-T. Nguyen、B. Le 和 M.-L. Nguyen，《用于越南语情感分析的多通道长短期记忆网络-卷积神经网络模型》，载于*2017*年第*9*届知识与系统工程国际会议（*KSE*），2017年，第24–29页。[26] N. X. Bach、D. V. Pham、N. D. Tai 和 T. M. Phuong，《面向情感分析的越南语比较句挖掘》，载于*2015*年第七届知识与系统工程国际会议（*KSE*），2015年，第162–167页。[27] H. T. Vo、H. C. Lam、D. D. Nguyen 和 N. H. Tuong，《越南语教育调查系统的主题分类与情感分析》，《亚洲计算机科学与信息技术期刊》，第6卷，第3期，第27–34页，2016年。[28] T. K. Tran 和 T. T. Phan，《从在线评论中挖掘观点目标和观点词》，《国际信息技术期刊》，第9卷，第3期，第239–249页，2017年9月。[29] Q.-T. Ha、T.-T. Vu、H.-T. Pham 和 C.-T. Luu，《越南语产品评论中基于特征的观点挖掘模型的升级》，载于主动媒体技术，N. Zhong、V. Callaghan、A. A. Ghorbani 和 B. Hu 编。柏林、海德堡：施普林格柏林海德堡，2011年，第173–185页。[30] T.-T. Vu、H.-T. Pham、C.-T. Luu 和 Q.-T. Ha，《越南语产品评论中的基于特征的观点挖掘模型》。柏林、海德堡：施普林格柏林海德堡，2011年，第23–33页。[31] D.-H. Phan 和 T.-D. Cao，《运用 Skip-Gram 词语估计和基于支持向量机的分类来挖掘越南语餐饮场所文本评论中的观点》，载于第五届信息与通信技术研讨会论文集，丛书 SoICT '14。美国纽约：计算机协会，2014年，第232–239页。[32] D.-H. Pham、A.-C. Le 和 T.-K.-C. Le，《为越南语评论学习语义表示以进行评分》，载于*2016*年第八届知识与系统工程国际会议（*KSE*），2016年，第193–198页。[33] V. D. Nguyen、K. V. Nguyen 和 N. L.-T. Nguyen，《用于情感分析的长短期记忆网络变体：越南语学生反馈语料库》，载于*2018*年第*10*届知识与系统工程国际会议（*KSE*），2018年，第306–311页。[34] T.-L. Truong、H.-L. Le 和 T.-P. Le-Dang，《将基于 BERT 的预训练语言模型用于越南语情感分析》，载于*2020*年第*7*届*NAFOSTED*信息与计算机科学会议（*NICS*），2020年，第362–367页。[35] K. V. Nguyen、V. D. Nguyen、P. X. V. Nguyen、T. T. H. Truong 和 N. L.-T. Nguyen，《UIT-VSFC：用于情感分析的越南语学生反馈语料库》，载于*2018*年第*10*届知识与系统工程国际会议（*KSE*），2018年，第19–24页。[36] V. A. Ho、D. H.-C. Nguyen、D. H. Nguyen、L. T.-V. Pham、D.-V. Nguyen、K. V. Nguyen 和 N. L.-T. Nguyen，《越南语社交媒体文本的情绪识别》，载于计算语言学，L.-M. Nguyen、X.-H. Phan、K. Hasida 和 S. Tojo 编。新加坡：Springer Singapore，2020年，第319–333页。[37] D. Van Thin、D.-V. Nguyen、K. Van Nguyen、N. L.-T. Nguyen，

和 A. H.-T. Nguyen，《越南语数据集上的方面与极性识别多任务学习》，载于计算语言学，L.-M. Nguyen、X.-H. Phan、K. Hasida 和 S. Tojo 编。新加坡：Springer Singapore，2020年，第169–180页。[38] H. D. Nguyen、T. Le、K. V. Tran、S. T. Luu、S. N. Hoang 和 H. T. Phan，《基于语法规则的产品评论多层级情感分析》，载于*SoMeT 2021*论文集，H. Fujita 和 H. Pérez-Meana 编，第337卷。IOS Press，2021年，第444–456页。[39] H. D. Nguyen、T. Huynh、S. N. Hoang、V. T. Pham 和 I. Zelinka，《基于语法结构和改进型自注意力网络的面向语言的情感分析》，载于第*15*届新型软件工程方法评估国际会议论文集，*ENASE 2020*，布拉格，捷克共和国，*2020*年*5*月*5*日至*6*日，R. Ali、H. Kaindl 和 L. A. Maciaszek 编。SCITEPRESS，2020年，第339–346页。[40] L. Luc Phan、P. Huynh Pham、K. Thi-Thanh Nguyen、S. Khai Huynh、T. Thi Nguyen、L. Thanh Nguyen、T. Van Huynh 和 K. Van Nguyen，《SA2SL：从基于方面的情感分析到面向商业智能的社交聆听系统》，载于知识科学、工程与管理，H. Qiu、C. Zhang、Z. Fei、M. Qiu 和 S.-Y. Kung 编。Cham：Springer International Publishing，2021年，第647–658页。[41] A. Joulin、E. Grave、P. Bojanowski 和 T. Mikolov，《高效文本分类的技巧集》，载于欧洲计算语言学协会分会第*15*届会议论文集：第*2*卷，短篇论文。计算语言学协会，2017年4月，第427–431页。[42] N. C. Lê、N. The Lam、S. H. Nguyen 和 D. Thanh Nguyen，《越南语情感分析：一种迁移学习方法》，载于*2020*年*RIVF*计算与通信技术国际会议（*RIVF*），2020年，第1–5页。[43] Q. T. Nguyen、T. L. Nguyen、N. H. Luong 和 Q. H. Ngo，《微调 BERT 以分析越南语评论的情感》，载于*2020*年第*7*届*NAFOSTED*信息与计算机科学会议（*NICS*），2020年，第302–307页。[44] O. T. Tran 和 V. T. Bui，《用于越南语基于方面的情感分析的 BERT 层次模型》，载于*2020*年第*12*届知识与系统工程国际会议（*KSE*），2020年，第269–274页。[45] J. Devlin、M.-W. Chang、K. Lee 和 K. Toutanova，《BERT：用于语言的深度双向变换器预训练

“理解”，载于《*2019*年计算语言学协会北美分会年会论文集：人类语言技术》，第*1*卷（长篇与短篇论文）。美国明尼苏达州明尼阿波利斯：计算语言学协会，2019年6月，第4171–4186页。[46] P. Bojanowski、E. Grave、A. Joulin和T. Mikolov，“利用子词信息丰富词向量”，计算语言学协会会刊，第5卷，146，2017年，第135页。–。[47] D. Q. Nguyen和A. Tuan Nguyen，“PhoBERT：越南语预训练语言模型”，载于计算语言学协会研究成果：*2020*年自然语言处理实证方法会议。线上：计算语言学协会，2020年11月，第1037–1042页。[48] R. Sennrich、B. Haddow和A. Birch，“使用子词单元翻译罕见词的神经机器翻译”，载于第*54*届计算语言学协会年会论文集（第*1*卷：长篇论文）。德国柏林：计算语言学协会，2016年8月，第1715–1725页。[49] A. Conneau、K. Khandelwal、N. Goyal、V. Chaudhary、G. Wenzek、F. Guzmán、E. Grave、M. Ott、L. Zettlemoyer和V. Stoyanov，“大规模无监督跨语言表征学习”，载于计算语言学协会第*58*届年会论文集。线上：计算语言学协会，2020年7月，第8440–8451页。[50] A. Gu、K. Goel和C. Ré，“利用结构化状态空间高效建模长序列”，载于国际学习表征会议（*ICLR*），2022年。[51] A. Baevski和M. Auli，“神经语言建模的自适应输入表征”，*CoRR*，卷号abs/1809.10853，2018年。[52] I. Loshchilov和F. Hutter，“修正Adam中的权重衰减正则化”，*CoRR*，卷号abs/1711.05101，2017年。[53] F. Pedregosa、G. Varoquaux、A. Gramfort、V. Michel、B. Thirion、O. Grisel、M. Blondel、P. Prettenhofer、R. Weiss、V. Dubourg、J. Vanderplas、A. Passos、D. Cournapeau、M. Brucher、M. Perrot和E. Duchesnay，“Scikit-learn：Python中的机器学习”，机器学习研究杂志，第12卷，2830，2011年，第2825页。–。[54] N. X. Bach、N. D. Linh和T. M. Phuong，“面向社交媒体文本的越南语词性标注”，载于神经信息处理，A. Hirose、S. Ozawa、K. Doya、K. Ikeda、M. Lee和D. Liu编。瑞士Cham：Springer International Publishing，2016年，第233–242页。