Computer Science and Application 计算机科学与应用, 2022, 12(2), 323-330
Published Online February 2022 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2022.122032

Hans汉斯

一种融合Text-Rank的旅游文本情感分析方法

张　举[1]，冯　翱[1]，张学磊[1]，刘　涛[1]，栗荣成[1]，周道华[2]，杨　陈[2]，曾　俊[2]

[1]成都信息工程大学计算机学院，四川　成都
[2]成都中科大旗软件股份有限公司，四川　成都

收稿日期：2022年1月15日；录用日期：2022年2月11日；发布日期：2022年2月18日

摘　要

对于在线平台和论坛中的旅游评论文本进行情感分析一方面有助于景区理解游客的需求，提高景区服务品质，另一方面为游客提供出游参考信息，以达到更好的旅游满意度，具有较高的应用价值。本文针对旅游评论文本普遍较长的特征，提出了融合Text-Rank的情感分类方法。在进行情感分类之前先使用Text-Rank方法对较长的文本进行自动摘要，使用摘要压缩后的内容作为输入进行情感分类。使用RNN、LSTM、Text-CNN、BERT等深度学习模型进行实验，结果显示融合后的方法在准确率等各项指标上均取得了一定程度的提升。该方法对于原始文本较长、包含多方面内容的输入信息具有较大价值，能够提高文本处理效率，提高分析精度。

关键词

旅游文本，情感分类，深度学习，Text-Rank，BERT

A Sentiment Analysis Method for Tourism Text Integrated with Text-Rank

Ju Zhang[1], Ao Feng[1], Xuelei Zhang[1], Tao Liu[1], Rongcheng Li[1], Daohua Zhou[2], Chen Yang[2], Jun Zeng[2]

[1]School of Computer Science, Chengdu University of Information Technology, Chengdu Sichuan
[2]DAQSoft (Chengdu) Co., Ltd., Chengdu Sichuan

Received: Jan. 15th, 2022; accepted: Feb. 11th, 2022; published: Feb. 18th, 2022

Abstract

Sentiment analysis of tourism comment text on online platforms and forums helps scenic spots to understand tourist needs and improve their service quality. It also provides valuable reference

文章引用: 张举, 冯翱, 张学磊, 刘涛, 栗荣成, 周道华, 杨陈, 曾俊. 一种融合 Text-Rank 的旅游文本情感分析方法[J]. 计算机科学与应用, 2022, 12(2): 323-330. DOI: 10.12677/csa.2022.122032

information for tourist satisfaction, with high value for both parties. Dominated by long text in tourism comment, a sentiment classification model integrated with Text-Rank is proposed. Before feeding original text into the classifier, the Text-Rank method generates a summary of the long text, and the compressed summary is used as the input for sentiment classification. RNN, LSTM, Text-CNN and Bert are used in the experiment. Its result shows that the fused method yields significant performance over the original model. The proposed method improves text processing efficiency and accuracy, especially when the original text is over certain length or contains multiple aspects.

Keywords

Tourism Text, Sentiment Classification, Deep Learning, Text-Rank, BERT

1. 引言

文本情感分析(Text Sentiment Analysis, TSA)是自然语言处理(Natural Language Processing, NLP)中的重要研究领域之一，其应用与发展前景良好，已被广泛应用于辅助决策、电子商务、心理治疗等多个领域[1]。近年来，随着互联网和移动互联网的飞速发展，旅游类产品迅速火热。因此，网络上产生了大量针对目的景点的主观评论。游客可以通过微博、携程、去哪儿、马蜂窝和同程等 APP 或者网站获取到其他人对各个景点的相关评论，以辅助他们进行旅游出行决策。通过分析旅游评论文本中的情感信息，我们可以掌握评论中所包含的情感倾向，帮助游客规避旅游出行风险，还可以给景点运营部门提供管理建议，具有较高的应用价值。

旅游评论的情感分类问题与传统的文本情感分类问题定义非常相似，但是如果直接套用传统的文本情感分类方法，无法取得很好的效果。传统情感分类模型主要是针对商品评论一类的短文本进行分析，其主要模型包括基于情感词典的方法、基于机器学习的方法以及基于深度学习的方法等[2]。而旅游评论文本由于其固有属性，大多数情况下都包含着大量的景点描述信息，因此其文本较长，天然就存在较多冗余和噪声，特征较为稀疏。因此，旅游评论文本的情感分析任务属于长文本分类问题，相比较于传统的情感分析而言，存在特征稀疏、难以充分提取等技术难点。

本文研究了如何解决旅游文本中的信息冗余、噪声以及特征稀疏的问题，提出了一种融合 Text-Rank 的旅游文本情感分类方法，并在 RNN、LSTM、Text-CNN、BERT 等深度学习模型上进行了实验。实验结果表明，在 BERT 基础上使用 Text-Rank 进行预处理，相比通常使用的直接截断方式分类模型准确率提高了 1.6%，在其他模型上也有不同幅度的提升。

2. 相关工作

早期的文本情感分析方法主要基于情感词典，但针对旅游评论这样的长文本，情感词典很难做到全面覆盖。而且由于旅游评论文本属于特定的产业领域，具有专业话术，存在着不少非通用的情绪表达语言，因此基于情感词典的方法具有较大的应用局限性。基于机器学习的方法具有更为广泛的适用性，情感分析领域常用的方法主要包括朴素贝叶斯、支持向量机和决策树等分类算法[3] [4] [5]。近年来，随着文本嵌入表征方式的广泛应用，以及大规模预训练模型的流行，基于深度学习的情感分类方法已经成为

了情感分析中的主流。

2.1. 基于情感词典的方法

基于情感词典的方法是文本情感分析中的基础模型，构建情感词典是这类方法中的关键工作[6]。情感词典的构建有两类，人工和自动构建。人工的方式首先对句子进行分词，给每个词语进行情感极性和强度的标注打分，然后汇总成情感词典，最后再通过查阅词典对待分析文本中的情感词进行加权求和，得到句子的情感极性。人工构建情感词典的方法有一个极大的缺陷，就是需要耗费大量的人力开销。并且这种方式在处理较为专业的文本时，由于参与构建情感词典的人员对该领域的熟悉程度和理解程度有差异，这将导致同一个词语在不同人的标注下情感极性的标注结果是不同的，甚至差距很大。自动构建情感词典的方法能较大地降低人工成本，该方法主要从特定领域的大量语料库中进行学习，实现对该领域的情感词进行提取。基于情感词典的方法由于词典的局限性，通常适用于特定领域的短文本分析，对于具有更大自由度的长文本，基于机器学习的方法和基于深度学习的方法将会有更好的表现。

2.2. 基于机器学习的方法

基于机器学习的方法将语料库中的数据分为训练数据和测试数据两部分，将文本表示为向量后使用训练数据构建分类模型，再使用训练好的分类模型对测试数据进行预测。分类的主要方法包括朴素贝叶斯、决策树、支持向量机等。基于机器学习的情感分析方法通常只是简单地根据文本中的关键词特征进行分类，无法理解文本中的语序和语义信息，模型泛化能力较弱。由于文本和语言表达的多样性，以及语言含义本身具有不断的发展演变的特点，基于机器学习的情感分析方法很难有效解决当前的各种分析任务[6] [7]。但由于基于机器学习的方法模型简单，对算力要求较低，因此也常在一些情感分类任务应用中作为辅助算法。

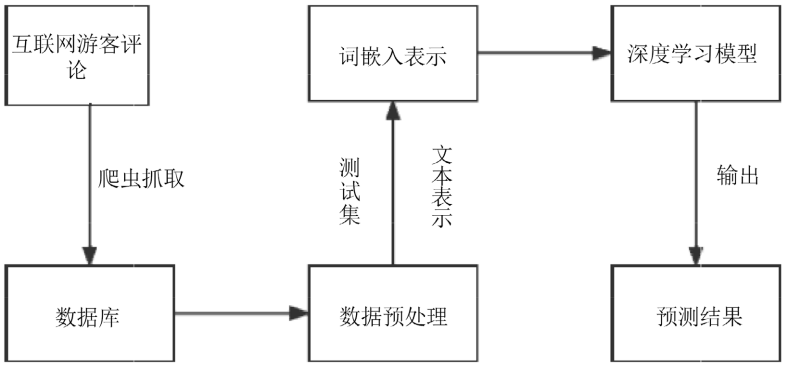

Figure 1. Flow chart of deep learning sentiment analysis of tourism comment text
图 1. 旅游评论文本的深度学习情感分析流程

2.3. 基于深度学习的方法

基于深度学习的情感分析方法流程如图 1 所示。随着 Word2Vec [8]和 Glove [9]的提出，使用预训练词向量的深度学习方法很快被应用于情感分析任务中。卷积神经网络(Convolution Neutral Network, CNN)因其能全方面地捕获提取文本信息和速度快的优点，常应用于长文本的情感分类任务中。循环神经网络(Recurrent Neural Networks, RNN)进行情感分析时，会随着文本的长度的增大提高模型训练成本，并且无法解决长距离依赖问题，因此大多应用于短文本情感分析任务中。长短时记忆网络(Long Short Term Memory, LSTM) [10]是一种特殊的 RNN，它能储存长距离信息，并捕获文本之间的依赖关系。为了有效

解决长文本分类时的特征稀疏问题，Yang 等提出了分层注意力网络(HAN) [11]，先分别对文档中的词语和句子进行编码以获取注意力权重，再将两部分权重整合为文档嵌入表达后加权平均进行长文本分类。在处理情感分类任务时，注意力机制能有效捕获文本中的情感词特征，因此常用做辅助模块，与其他深度学习方法结合使用。

基于词向量的深度学习方法非常依赖词向量的质量。一方面，训练阶段需要大量的语料库和极高的计算成本才能得到表征能力强的词向量；另一方面，针对特定的领域，通用的词向量往往不能有很好的表征能力。因此，基于词向量的深度学习方法在特定的领域中要想取得好的结果，较为理想的方案就是使用该领域的大规模语料库，以训练适用于这个领域的词向量。为应对特定领域语料不足的现实情况，研究者提出使用大规模通用语料库训练出一个预训练语言模型，再根据特定的领域，对这个预训练语言模型进行微调，从而适用于多个领域。BERT [12]及其相关变体是近年来最为流行的大规模预训练语言模型，广泛应用于短文本任务中。但由于模型参数的限制，BERT 对于输入数据的长度有一定的要求，如果输入长度超过阈值，会直接将句子截断，而在被去掉的内容中，可能还存在有价值的语义信息，BERT 实际上忽略了这些信息。因此，BERT 对于长文本情感分类任务存在明显的信息损失。

3. 模型

本节主要介绍旅游文本情感分析任务，并提出一种融合 Text-Rank [13]的旅游文本情感分析方法，用摘要提取的方法对长文本进行有效情感分析。

3.1. 任务描述

给定一条总词语数量为 n 的句子，其文字表示为 $T=(w_1,w_2,\cdots,w_n)$，通过模型进行分类判断，获得句子 T 的情感极性 $S=(\text{Positive},\text{Negative})$，其中 Positive 代表正面情感，Negative 代表负面情感。

3.2. 方法流程

本文提出了一种融合 Text-Rank 的旅游文本情感分析方法，用于处理基于旅游评论的长文本，其整体流程如图 2 所示，包括输入部分，文本摘要部分、词嵌入部分、深度学习模型部分以及情感分类部分。

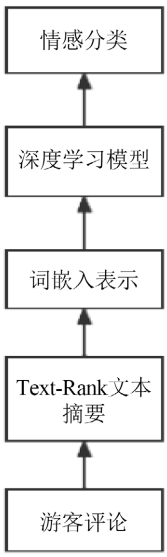

Figure 2. Flow chart of classification model's data processing
图 2. 分类模型的数据处理流程

3.3. Text-Rank

下面对其中使用 Text-Rank 进行文本摘要的方法进行详细说明。

为了保证输入模型的句子长度均保持在短文本的范围，我们引入了 Text-Rank 方法，对输入的文本数据进行预处理。对于长度超过 300 个字的句子，我们先对其求摘要，然后以摘要文本替换原文本作为后续深度学习模型的输入。

Text-Rank 算法进行文本摘要的步骤如下：

1) 对文本进行断句，分割。
2) 对每个句子进行分词并过滤掉停用词。
3) 计算 BM25 相关性矩阵。
4) 迭代投票。
5) 排序输出结果，通常取前三句拼接后作为摘要。

Text-Rank 算法的公式如下：

$$WS\left(V_i\right)=\left(1-d\right)+d*\sum_{V_j\in In\left(v_i\right)}\left(W_{ji}\Big/\sum_{V_k\in Out\left(V_j\right)}W_{jk}*WS\left(V_j\right)\right)$$

等式左边表示一个句子的权重(weight_sum, WS)，右侧的求和表示每个相邻句子对本句子的贡献程度。V_i 表示句子 i，d 表示 damping factor，用于做平滑。求和的分母 W_{ij} 表示句子 i 和句子 j 的相似程度，分母又是一个 weight_sum，而 $WS(V_j)$代表上次迭代句子 j 的权重。整个公式是一个迭代的过程。W_{ij} 使用 BM25 算法来计算两个句子之间的相似度，其主要思想是：对 Query 进行语素解析，生成语素 q_i；然后，对于每个搜索结果 D，计算每个语素 q_i 与 D 的相关性得分，最后，将 q_i 相对于 D 的相关性得分进行加权求和，从而得到 Query 与 D 的相关性得分。其一般性公式如下：

$$\text{Score}\left(Q,d\right)=\sum_{i}^{n}W_i*R\left(q_i,d\right)$$

其中，Q 表示 Query，q_i 表示 Q 解析之后的一个语素（对中文而言，我们可以把对 Query 的分词看作为语素分析，每个词看成语素）；d 表示一个搜索结果文档；W_i 表示语素的权重，使用 TF-IDF 进行计算；$R\left(q_i,d\right)$ 表示语素 q_i 与文档 d 的相关性得分。

4. 实验与结果分析

4.1. 数据来源及评估指标

实验所用的数据集来自于各大旅游网站和 APP，包括携程、去哪儿、同城、蚂蚁窝和微博等。我们采用网络爬虫的方式共采集 200,000 余条旅游评论语料，经过人工筛选后，保留 12,000 条包含情感极性的评论，通过人工标注的方式对这些评论进行情感标记，其中正面评价 6000 条，负面评价 6000 条。我们将这 12,000 条旅游评论文本按各类比例分为 10,000 条训练数据、1000 条验证数据和 1000 条测试数据。对数据集中数据按照文本长度(超过 300 与不超过 300)进行分类统计，结果如图 3 所示，可以看到句子长度超过 300 的句子大概占整个数据集的 10%左右。

我们使用训练数据学习模型参数，基于验证集上的各项指标优化超参，并使用验证集上表现最好的模型进行测试。

根据样本真实情感与模型预测情感可将测试数据的结果划分为正确正例(TP)、错误正例(FP)、正确负例(TN)、错误负例(FN)四类。采用分类精确度 Accuracy 对模型学习结果进行评估。计算公式为：

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$$

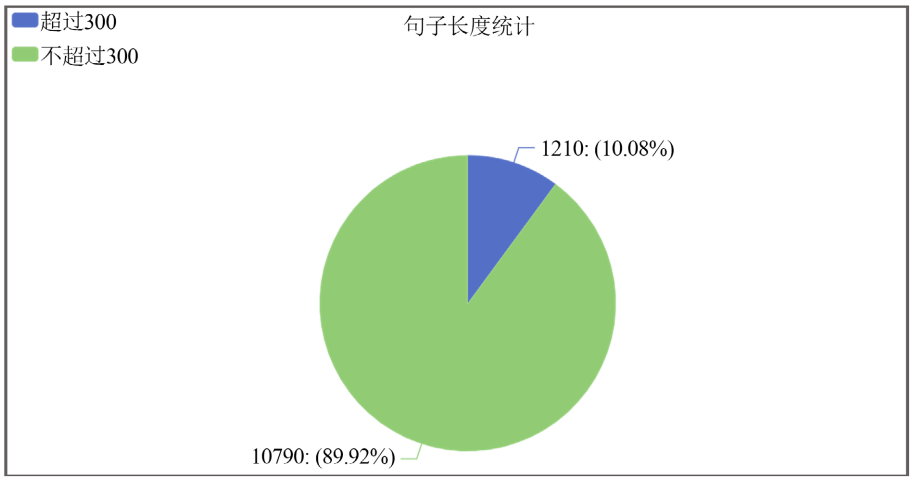

Figure 3. Statistical chart of Sentence length
图 3. 句子长度统计图

4.2. 实验环境

实验使用的深度学习框架是 Pytorch [14]，在单张 Nvidia RTX 3090 GPU 上进行训练，训练参数如表 1 所示。

Table 1. Model parameters in training phase
表 1. 模型训练参数

参数名	参数值
Batch size	512
Learning rate	5E−5
Optimizer	Adam

4.3. 模型对比

为了全面评价本文提出模型的性能，我们在上述的数据集上进行了情感分类实验，并与多种基线模型进行对比。实验结果如表 2 所示，在 RNN、LSTM、Text-CNN [15]和 BERT 四个基准模型基础上，融合 Text-Rank 的文本摘要方法后分类精度均有一定程度的提高。表中标为原始的指标代表输入文本未经过任何处理，作为基准模型，标为融合的指标代表融合了 Text-Rank 之后的方法，作为对照模型。

Table 2. Performance of different models on experiment dataset
表 2. 不同模型在实验数据集上的性能表现

模型	Accuracy (原始)	Accuracy (融合)
RNN	76.50	**76.90**
LSTM	82.15	**83.10**
Text-CNN	85.50	**86.20**
BERT	91.75	**93.35**

1) RNN：基于旅游评论文本数据训练的 300 维 Word2Vec 中文词向量表示评论文本，将它们输入到隐藏层，经过两个隐藏层，每个隐藏层有 128 个隐藏单元，多层循环迭代之后，输入到线性层，从而对旅游文本进行情感分类。

2) LSTM：使用基于旅游评论文本数据的 300 维 Word2Vec 中文词向量表示评论文本，再通过一层双向 LSTM，两个隐藏层，每个隐藏层有 64 个隐藏单元去提取文本特征，使用提取的特征向量输入到线性层对旅游文本进行情感分类。

3) Text-CNN：基于旅游评论文本数据训练的 300 维 Word2Vec 中文词向量表示评论文本，并使用 3 个卷积核，卷积核的大小分别为 2，3，4 提取文本特征，然后将它们输入到线性变换层，从而对旅游文本进行情感分类。

4) BERT：基于旅游评论文本数据训练的 300 维 token embeddings 中文词向量表示评论文本，将它们输入到 BERT_BASE 模型，从而对旅游文本进行情感分类。

4.4. 结果分析

从表 2 的实验结果可以看出，使用 Text-Rank 进行输入文本的预处理，在处理旅游评论这类长文本较多的情感分类任务时相比于各个基准模型具有更好的表现。融合了 Text-Rank 的方法相对于原始的 Text-CNN、LSTM、RNN、BERT 平均准确率分别提升了 0.7%、0.95%、0.4%、1.6%，其中对于 BERT 这个原始精度最高的基准模型进一步取得了显著的提升。其主要原因是 BERT 对于输入文本有长度限制，对于过长的输入直接进行了截断操作，导致明显的信息损失。在长文本的情感分类任务中，融合了文本摘要算法之后能够更好地保留原始文本的语义信息，提升情感分析模型理解文本的能力，以达到更好的分类效果。

5. 结束语

传统的深度学习网络在原始文本过长时会进行裁剪操作，以保证处理模型的参数统一，其后果是丢失了部分原文中的所含有的语义信息，但该部分很多时候会包含影响分类决策的关键因素，从而导致分类的准确率下降。

目前大多数情感分类工作都是面向以网络商品评论为主的短文本，由于句子短、特征少，较为简单的深度学习模型就能取得较好效果。旅游评论文本中既包含短文本，又有长文本，因此旅游评论的情感分析任务的主要问题在于部分篇幅较长且包含复杂特征，简单的截断方法有可能损失原始文本中的关键性信息。本文提出了一种融合 Text-Rank 的旅游文本情感分析方法，通过文本摘要算法，将数据集中的长文本在语义信息损失较小的前提下转为了短文本，能够较为有效地克服长文本情感分析中的主要技术障碍。本文所提出的方法相比于传统的深度学习方法有明显的性能优势，分类准确率有较大提升。

词语与文档之间相关性权重和词语与词语之间相关性权重的计算方式有很多种，本文使用较为简单的 TF-IDF 和 BM25 计算，取得了一定的性能提升。在后续的工作中，将尝试使用表征能力更强的计算方式，如使用互信息法(Mutual Information, MI) [16]或信息增益法(Information Gain, IG) [17]计算词语与文档的相关性，使用情感倾向 SO-PMI (Semantic Orientation Point-wise Mutual Information, SO-PMI) [18]计算词语与词语的相关性等，而对于领域特征的有效识别和提取会成为模型上限提升的主要手段。

基金项目

四川省科技计划资助(立项编号：2020YFG0168)。

参考文献

[1] 王颖洁, 朱久祺, 汪祖民, 白凤波, 弓箭. 自然语言处理在情感分析领域应用综述[J/OL]. 计算机应用, 1-12. http://kns.cnki.net/kcms/detail/51.1307.TP.20210928.1611.014.html, 2021-12-27.

[2] 王婷, 杨文忠. 文本情感分析方法研究综述[J]. 计算机工程与应用, 2021, 57(12): 11-24.

[3] Suykens, J.A.K. and Vandewalle, J. (1999) Least Squares Support Vector Machine Classifiers. *Neural Processing Letters*, **9**, 293-300. https://doi.org/10.1023/A:1018628609742

[4] Safavian, S.R. and Landgrebe, D. (1991) A Survey of Decision Tree Classifier Methodology. *IEEE Transactions on Systems, Man, and Cybernetics*, **21**, 660-674. https://doi.org/10.1109/21.97458

[5] Rish, I. (2001) An Empirical Study of the Naive Bayes Classifier. *IJCAI* 2001 *Workshop on Empirical Methods in Artificial Intelligence*, 4-10 August 2001, 41-46.

[6] 刘爽, 赵景秀, 杨红亚, 徐冠华. 文本情感分析综述[J]. 软件导刊, 2018, 17(6): 1-4+21.

[7] 夏海峰, 陈军华. 基于文本挖掘的投诉热点智能分类[J]. 上海师范大学学报(自然科学版), 2013, 42(5): 470-475.

[8] Mikolov, T., *et al*. (2013) Efficient Estimation of Word Representations in Vector Space.

[9] Pennington, J., *et al*. (2014) Glove: Global Vectors for Word Representation. *Proceedings of the* 2014 *Conference on Empirical Methods in Natural Language Processing*, Doha, 25-29 October 2014, 1532-1543. https://doi.org/10.3115/v1/D14-1162

[10] Hochreiter, S. and Schmidhuber, J. (1997) Long Short-Term Memory. *Neural Computation*, **9**, 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

[11] Yang, Z.C., *et al*. (2016) Hierarchical Attention Networks for Document Classification. *Proceedings of the* 2016 *Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, San Diego, June 2016, 1480-1489. https://doi.org/10.18653/v1/N16-1174

[12] Devlin, J., *et al*. (2018) Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding.

[13] Mihalcea, R. and Tarau, P. (2004) Textrank: Bringing Order into Text. *Proceedings of the* 2004 *Conference on Empirical Methods in Natural Language Processing*, Barcelona, 25-26 July 2004, 404-411.

[14] Paszke, A., Gross, S., Massa, F., *et al*. (2019) Pytorch: An Imperative Style, High-Performance Deep Learning Library. *Advances in Neural Information Processing Systems*, **32**, 8026-8037.

[15] Kim, Y. (2014) Convolutional Neural Networks for Sentence Classification. *Proceedings of the* 2014 *Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), Doha, October 2014, 1746-1751. https://doi.org/10.3115/v1/D14-1181

[16] Viola, P. and Wells III, W.M. (1997) Alignment by Maximization of Mutual Information. *International Journal of Computer Vision*, **24**, 137-154. https://doi.org/10.1023/A:1007958904918

[17] Kent, J.T. (1983) Information Gain and a General Measure of Correlation. *Biometrika*, **70**, 163-173. https://doi.org/10.1093/biomet/70.1.163

[18] Turney, P.D. and Littman, M.L. (2002) Unsupervised Learning of Semantic Orientation from a Hundred-Billion-Word Corpus.