

# 基于 TextCNN 和 LightGBM 的导游违规行为检测

刘昌澍<sup>1,2</sup>, 李响<sup>2,3</sup>, 詹瑾瑜<sup>1,2</sup>, 江维<sup>1</sup>, 李博智<sup>1</sup>, 曹扬<sup>2,3</sup>, 杨瑞<sup>2,3</sup>

(1. 电子科技大学 信息与软件工程学院, 四川 成都 610054;

2. 中电科大数据研究院有限公司, 贵州 贵阳 550022;

3. 提升政府治理能力大数据应用技术国家工程实验室, 贵州 贵阳 550022)

**摘要:**人工处理旅游评论需要耗费大量人力,如何自动分析旅游评论检测出导游违规行为,为旅游监管提供依据,成为一个迫切需要解决的热点问题。该文提出了一种基于 TextCNN 和 LightGBM 的导游违规行为检测方法,首先构建旅游评论的文本卷积神经网络(TextCNN)从海量旅游评论信息中筛选出负面评论;再将这些负面评论送入基于梯度提升决策树(LightGBM)的导游违规行为分类模型,分析得到导游违规行为的具体类型分类及分类概率。使用准确率、召回率、F1 值等多个性能指标对提出的模型进行测试与分析,实验数据表明,基于 TextCNN 和 LightGBM 的导游违规行为检测方法比一些其他主流方法和模型(SVM、LSTM、XGBoost 等)具有更好的准确性和合理性。同时,该方法应用在实际旅游大数据系统中可以得到 91.57% 的准确率。

**关键词:**自然语言处理;情感分析;导游违规行为;文本卷积神经网络;梯度提升决策树

中图分类号:TP391

文献标识码:A

文章编号:1673-629X(2021)05-0143-07

doi:10.3969/j.issn.1673-629X.2021.05.025

## Illegal Tour Guide Behavior Detection Based on TextCNN and LightGBM

LIU Chang-shu<sup>1,2</sup>, LI Xiang<sup>2,3</sup>, ZHAN Jin-yu<sup>1,2</sup>, JIANG Wei<sup>1</sup>,

LI Bo-zhi<sup>1</sup>, CAO Yang<sup>2,3</sup>, YANG Rui<sup>2,3</sup>

(1. School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu 610054, China;

2. CETC Big Data Research Institute Co., Ltd., Guiyang 550022, China;

3. Big Data Application on Improving Government Governance Capabilities National Engineering Laboratory, Guiyang 550022, China)

**Abstract:** It takes a lot of manpower to process tourist comments manually. How to process tourist comments automatically and provide the evidence for tourism regulation has become an issue that needs to be addressed urgently. We propose an illegal tour guide behavior detection method based on TextCNN and LightGBM. Firstly, the tour guide negative comment identification model based on text convolutional neural networks (TextCNN) is presented to filter the tourist comments by emotion analysis. Then the filtered negative tourist comments are sent into the illegal tour guide behavior detection model based on light gradient boosting machine (LightGBM) to obtain the corresponding classification. We use precision, recall, F1-score to evaluate the performance of the proposed models. The experiment shows that the proposed illegal tour guide behavior detection method has better accuracy and rationality compared with other models including SVM, LSTM, XGBoost, etc. And the proposed method can achieve 91.57% precision in the practical tourism big data system.

**Key words:** natural language processing; emotion analysis; illegal tour guide behaviors; TextCNN; LightGBM

## 0 引言

随着人民生活水平的提高,国内旅游产业蓬勃发

展。2018 年国内旅游人数 55.39 亿人次,比上年同期增长 10.8%。初步测算,全年全国旅游业对 GDP 的综

收稿日期:2020-06-12

修回日期:2020-10-12

**基金项目:**提升政府治理能力大数据应用技术国家工程实验室开放基金项目(W-2019007);四川省科技计划项目(2018CC0136);中科院计算机体系结构国家重点实验室开放课题(CARCH201811);中央高校基本科研业务费(ZYGX2018J077, ZYGX2019J078)

**作者简介:**刘昌澍(1998-),男,CCF 会员(B7774G),研究方向为软件工程与自然语言处理;通信作者:詹瑾瑜(1978-),女,博士,副教授,CCF 会员(B3099M),研究方向为深度学习、大数据处理。

合贡献为 9.94 万亿元,占 GDP 总量的 11.04%<sup>[1]</sup>。但遗憾的是,由于利益的驱动、市场监管不足、导游职业素质较低等原因,近些年频频出现不合理低价竞争、随意加点和强制消费等导游违规现象,负面报道层出不穷,严重影响了导游的社会形象与诚信度<sup>[2]</sup>。近年来,随着网络技术的发展,旅游业开始和互联网产业结合,出现了像去哪儿、途牛、携程等在线旅游平台,产生大量如评论、推荐等文本数据。这些文本数据中包含对于导游违规行为的负面评论或投诉。如何更好地利用这些数据实现对导游行为的监管,进一步完善规范旅游市场,一直是有关部门关注的话题。旅游评论等文本信息可以利用自然语言处理相关技术进行文本信息挖掘,并有大量范例可供参考借鉴。Kamran Kowsari 等人<sup>[3]</sup>对经典文本分类任务的流程定义,介绍了文本预处理和文本特征抽取的方法并对包括 SVM、KNN、决策树等机器学习算法进行比较。Zhang Lei 等人<sup>[4]</sup>在文本分类的基础上进一步挖掘情感分析问题,并比较了包括 CNN、RNN、LSTM 和 HAN 等基于深度学习的 state of art 模型。同时,文本分类或情感分析可以用来处理负面评论或投诉信息,在中文语境下也有一些实际应用案例。例如,梁昕露等人<sup>[5]</sup>针对电信行业的客户投诉系统,提出了使用向量空间模型做文本特征,用 SVM 做分类器的分类方法,在 13 万条测试数据上取得了 70% 以上的准确率。但是该方法存在准确率在分类上并不均衡的问题。余本功等人<sup>[6]</sup>使用 BTM 和 Doc2vec 模型构建较低的 SVM 输入空间并引入了集成学习的思想,提出了基于 nB-SVM 的投诉识别方法,一定程度上改善了由于数据不均衡引发的问题。近年来随着深度学习领域的快速发展,出现了很多深度学习方法在中文文本分类及投诉处理的实践。例如,郑诚等人<sup>[7]</sup>等人提出了 BLSTM\_MLPCNN 神经

网络模型,并将字符级向量联合词向量作为 BLSTM 的输入,在 5 个英文标准数据集上进行实验中均取得了较好的效果。万圣贤等人<sup>[8]</sup>提出了包括 MaxBiLSTM 和 ConvBiLSTM 的双向 LSTM 模型,更加高效地提取中间文本特征,该方法在公开数据集上取得了较好的效果。段立等人<sup>[9]</sup>针对 95598 客服投诉工单,提出了基于 XGBoost 的分类归档的方法,准确率在 83% ~ 91% 左右,有较好的效果。文献[10]提出基于 word2vec 和双向 LSTM 的情感分类深度模型,解决社交网络文本传统情感分类模型存在先验知识依赖以及语义理解不足的问题。

该文主要研究导游违规行为的识别检测问题,首先构建了基于文本卷积神经网络(TextCNN)的旅游评论情感分析模型,采用预先训练的旅游话题词向量表示整段文本,使用多种尺寸的卷积核捕捉文本序列中上下文之间的关联,通过池化、全连接等操作得到文本对应的特征表达并判断情感倾向,达到从旅游评论中精准识别出导游相关的负面评论的目的。然后提出了基于轻量级梯度提升决策树(LightGBM)的导游违规行为分类模型,利用提升(boosting)的方法组合决策树,在训练过程中根据梯度差异给予不同样本不同的权重,实现对导游违规行为的准确分类,为旅游监管提供参考依据。进行了多组对比实验评估模型性能,实验结果表明,提出的方法可以有效从大量旅游评论中检测出导游违规行为及其类别,准确率达到 91.57%。

## 1 系统模型

基于 TextCNN 和 LightGBM 的导游违规行为检测系统的框架如图 1 所示,可以处理来自在线旅游平台的文本信息。

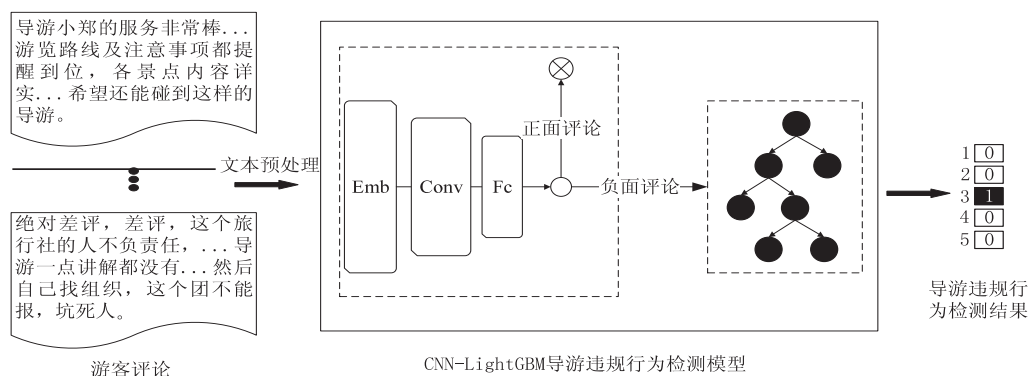


图 1 导游违规行为检测系统框架

首先对来自游客的导游相关评论进行文本预处理,包括分词以及去停用词。然后将进行词嵌入,将文本转化为向量,构建模型的输入。再使用 TextCNN 对输入文本的特征进行抽取与分类,根据情感倾向,从导

游相关评论中识别出负面评论。最后使用基于 Boosting 方法的 LightGBM 分类框架,对导游相关的负面评论进行违规行为检测与分类。导游违规行为有以下 5 类:(1)强迫消费;(2)殴打辱骂游客;(3)擅自

更改行程;(4)餐饮/住宿条件与合同不符;(5)不具备相关从业资格。

## 2 基于 TextCNN 和 LightGBM 的导游违规行为检测方法

### 2.1 基于 TextCNN 的导游负面评论识别

#### 2.1.1 TextCNN 建模

TextCNN<sup>[11]</sup>采用多个尺寸不同的卷积核来提取文本信息,能够更好地捕捉到上下文之间的关联。该文本构建的基于 TextCNN 的导游负面评论识别模型如图 2 所示,主要包括嵌入层、卷积层、池化层、全连接层和输出层<sup>[12]</sup>。

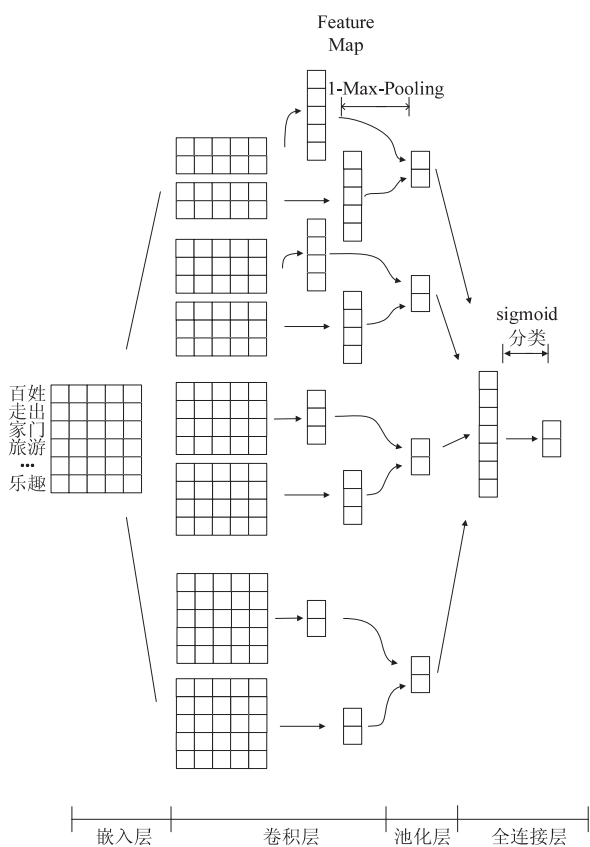


图 2 基于 TextCNN 的导游负面评论识别模型

#### (1) 嵌入层。

嵌入层(输入层)将一段文本转换为一个  $\text{maxlen} \times \text{dim}$  的二维矩阵,  $\text{maxlen}$  指输入文本可包含的词语数量最大值。该文统计去掉停用词之后的游客评论文本序列的长度,共计 76% 长度在 50 以内,17.6% 在 50 和 100 之间,5.8% 在 100 之上。因此选定  $\text{maxlen}$  为 100, 长度不足的进行补齐,长度超过的进行截取。

设  $x_i$  为某一个词对应的长度为  $\text{dim}$  的 word2vec<sup>[13-14]</sup> 词向量,通过将文本序列中每个词对应的词向量连接起来,就可以得到整个文本序列的词向量表达矩阵:

$$X = x_1 \oplus x_2 \oplus \dots \oplus x_{\text{maxlen}} \quad (1)$$

#### (2) 卷积层。

在卷积层对上一层的输出进行卷积运算,得到多个尺寸不同的特征图 (Feature Map)。卷积运算的过程可以表达为:

$$c_j = f\left(\sum_{j=i}^{i+h-1} x_j \cdot W + b\right) \quad (2)$$

其中,  $c_j$  指卷积运算得到的一个特征,  $b$  为偏置,  $W$  为卷积核矩阵,  $f$  为一个非线性函数。多个尺寸不同的卷积核在游客评论文本上形成多个跨度不同的滑动窗口,用来计算各个窗口内的单词之间的联系,如图 3 所示。

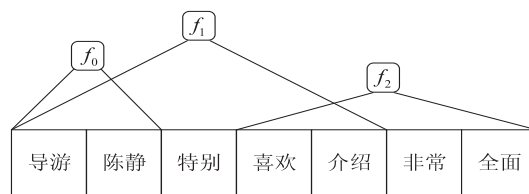


图 3 提取旅游评论特征的滑动窗口

经过卷积层一个高度为  $h$  的卷积核产生一个特征图 (feature map)  $C$ :

$$C = [c_1, c_2, \dots, c_{\text{maxlen}}] \quad (3)$$

#### (3) 池化层与全连接层。

在池化层,将特征图作为输入,进行维数降低。该文使用 1-max-pooling 的方式处理特征图,从中选取最大值。全连接层将池化结果拼接起来,从而得到了一段文本的特征。

#### (4) 输出层。

通过 sigmoid 函数得到导游评论中游客各种情感倾向的概率,并在输出层输出:

$$\text{sigmoid}(x) = \frac{1}{1 + e^{-x}} \quad (4)$$

构建深度学习模型之后,首先设定模型的损失函数为 binary\_crossentropy。

$$L_{bc} = \sum_{i=0}^n (y_i \times \log y'_i) + (1 - y_i) \times \log(1 - y'_i) \quad (5)$$

其中,  $y_i$  是真实值,  $y'_i$  是预测值。

#### 2.1.2 TextCNN 参数设置

该文使用 Keras 框架搭建神经网络。设定  $\text{dim}$  为 300 并在嵌入层使用预先训练的旅游话题词向量做权重。

在卷积层,设置了 4 种不同高度的卷积核(2,3,4,5),每个卷积核的宽度和词向量的  $\text{dim}$  长度相等。每种卷积核各自设置 100 个 filter,设定非线性函数为 Relu 函数,将分别输出  $99 \times 1$ ,  $98 \times 1$ ,  $97 \times 1$ ,  $96 \times 1$  的四种特征图。在全连接层使用 dropout 和  $l_2$  正则化的方式抑制训练过程中的过拟合程度。

## 2.2 基于 LightGBM 的导游违规行为检测

### 2.2.1 LightGBM 建模

文中数据集本身存在数据不平衡的问题。采用集成学习通过集成元分类器分类结果的方式提升模型的泛化能力,进而提升分类器的性能<sup>[15]</sup>。LightGBM<sup>[16]</sup>采用基于 Boosting 方法的分类框架,可以更好地适应不平衡数据。

该文采用简单数据增强 EDA<sup>[17]</sup>的方式改善数据分布。具体措施包括:同义词替换、删除个别词语、交换词语在句子中顺序、随机插入新词等。

#### (1) 导游投诉文本向量化。

首先进行中文分词和文本预处理(去停用词),可以将一段短文本  $S$  处理成一段长度为  $n$  词语序列:

$$S = s_1 + s_2 + \cdots + s_n \quad (6)$$

令  $v_i$  为  $s_i$  在旅游话题词向量中对应的词向量,则可以计算得到短文本  $S$  的向量化表示  $V$ :

$$V = \frac{\sum_{i=1}^n v_i}{n} \quad (7)$$

#### (2) 计算梯度与生成决策树。

LightGBM 组合多个回归决策树  $T$  以得到最终的结果<sup>[18]</sup>:

$$f_T(x) = \sum_{i=1}^T f_i(x) \quad (8)$$

其中,  $f_i(x)$  指单个决策树模型。LightGBM 训练过程中添加决策树时第  $t$  步的损失可以表示成如下形式:

$$L_t = \sum_{i=1}^n L(y_i, F_{t-1}(x_i) + f_t(x_i)) \quad (9)$$

当损失函数  $L$  为均方函数时,采用损失函数的负梯度作为残差  $r$  的近似值:

$$r = \frac{\partial L(y, f(x_i))}{\partial f(x_i)} \quad (10)$$

导游违规行为检测的决策树生成算法如算法 1 所示:

算法 1: 导游违规行为检测决策树生成算法。

输入: 模型训练数据  $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$

输入: 树的最大深度  $d$

1: best\_split\_point  $\leftarrow 0$

2:  $y_i^{(0)} \leftarrow 0$

3:  $t \leftarrow 0$

4: 设定目标函数:  $\text{obj}(\theta) = \sum_{i=1}^n L(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k)$

5: while(  $t < d$  and 未满足迭代停止的条件) do:

6: best\_split\_point  $\leftarrow \text{find\_best\_split}()$

7: 在 best\_split\_point 根据 leaf\_wise 策略分裂决策树

8: 更新训练结果  $y_i^t = y_i^{t-1} + f_t(x_i)$

9:  $t++$

10: end while

其中, 首先输入导游违规行为数据集和树的最大深度; 在第 4 行设定导游违规行为检测树的目标函数, 包括对违规行为进行预测产生的损失与树的复杂度。在第 6~9 行计算目标函数, 确定梯度下降的方向, 计算最优切分点 best\_split\_point 并按照 leaf\_wise 策略分割导游违规行为检测决策树并完成模型的更新。

#### (3) 减少采样与划分最优节点。

LightGBM 使用单边梯度下降算法 (gradient-based one side sampling, GOSS), 减少计算量。GOSS 使用梯度作为样本权重, 重点关注梯度较大的导游违规行为样本<sup>[19]</sup>:

$$G = \sum_{x \in A} g_i + \frac{1-a}{b} \sum_{x \in B} g_i \quad (11)$$

其中,  $A$  指大梯度负面评论样本;  $B$  指小梯度负面评论样本;  $\frac{1-a}{b}$  为小梯度样本采样系数。之后遍历导游违规行为样本并计算信息增益, 在节点  $d$  上特征  $j$  的信息增益计算公式为<sup>[16]</sup>:

$$V_j(d) = \frac{1}{n} \left[ \frac{(\sum_{x_i \in A_d} g_i + \frac{1-a}{b} \sum_{x_i \in B_d} g_i)^2}{n_d^j(g)} + \frac{(\sum_{x_i \in A_r} g_i + \frac{1-a}{b} \sum_{x_i \in B_r} g_i)^2}{n_r^j(g)} \right] \quad (12)$$

其中,  $n$  为使用样本总数,  $g_i$  为损失,  $A_l$ 、 $A_r$ 、 $B_l$  和  $B_r$  为划分在节点左右的样本。

在减少采样的基础上计算信息增益并划分最优节点的算法如算法 2 所示:

算法 2: 最优节点划分算法 find\_best\_split。

输入: 训练数据  $X, Y$

输入: 大梯度样本采样率  $a$ , 小梯度样本采样率  $b$

1: gain\_list  $\leftarrow \{\}$

2: sorted  $\leftarrow$  降序排列( $X$ )

3: topN  $\leftarrow a \times \text{len}(X)$

4: rand( $N$ )  $\leftarrow b \times \text{len}(X)$

5: fact  $\leftarrow \frac{1-a}{b}$

6: top\_set  $\leftarrow \text{sorted}[0:\text{topN}]$

7: rand\_set  $\leftarrow$  随机选取( $\text{sorted}[\text{topN}:\text{len}(X)]$ , rand( $N$ ))

8: used\_set  $\leftarrow \text{top\_set} + \text{rand\_set}$

9: for  $j$  in  $X$ . features:

10: gain  $\leftarrow$  增益计算( $j$ , used\_set)

11: gain\_list.append(gain)

12: end for

13: gain\_max  $\leftarrow \max(\text{gain\_list})$

14: ind  $\leftarrow \text{gain\_list.index(gain\_max)}$

15: return ind, gain\_max

其中, 首先输入导游违规行为训练数据集和样本采样率; 第 1~8 行先按照梯度降序排列导游违规行为

样本,再根据大梯度样本采样率  $a$  和小梯度样本采样率  $b$  重新组合需要遍历的样本,从而减少样本数量,第 9~14 行计算导游违规行为样本各个属性上的信息增益,并得到最大信息增益。

### 2.2.2 LightGBM 建模

该文使用 Python3.7 环境下的 LightGBM API 进行模型的搭建和训练。模型的超参数将会影响模型预测的准确率。为了取得更好的效果,在训练过程中不断对超参数进行寻优,最后得到的一组较好参数如表 1 所示。

表 1 LightGBM 主要超参数

参数	描述	取值
max_depth	决策树的最大深度,防止过拟合	6
min_data_in_leaf	叶子可能具有的最小记录数,防止过拟合	101
feature_fraction	每轮训练随机选择用来建立决策树的特征比例	0.6
bagging_fraction	每轮训练使用数据比例	0.6
lambda_l1	正则化参数,控制过拟合	1.0
lambda_l2	正则化参数,控制过拟合	2.0

## 3 实验分析

该文设计并实现了基于 TextCNN 和 LightGBM 的导游违规行为检测模型,能够完成对旅游评论的负面评论识别和导游违规行为检测。进行了多组实验评估所提出的模型和方法,并与其他主流算法和模型进行了比较。

### 3.1 实验设置和数据来源

实验的硬件是:CPU 为 Intel i7 9700K,内存为 16G RDD4,显卡为两块 Nvidia RTX 2080ti,运行环境为 Linux 操作系统(Ubuntu 16.04.6)。全部实验代码由 Python3.7 编写。神经网络由 Keras 框架搭建,使用 TensorFlow 作为框架后端。

当前关于导游违规行为的识别暂无公开数据集,该文使用爬虫构建数据集,使用的实验数据来源有:(1)去哪儿网获取的成都、北京、上海和广州等地的景点评论信息;(2)人民网 315 旅游投诉平台的投诉信息。并对这些实验文本进行数据清洗,筛选出和导游相关的评论信息。

### 3.2 模型性能评估

使用准确率(precision)、召回率(recall)、F1-score 和正确率(accuracy)等指标评估模型性能。

#### 3.2.1 导游负面评论识别模型评估与分析

选取获取的共 13 000 条评论进行实验,包括好评(正面)7 000 余条,差评(负面)6 000 余条,如表 3 所示。按照 0.8 的比例划分训练集和测试集。在训练

集中,划分出 20% 的数据做验证集。

数据集的预处理使用 Keras 自带的 Tokenizer 转化为序列。为了抑制过拟合程度加入 dropout 层(设置值为 0.5)并添加 l2 正则化(值为 0.001)。

基于 TextCNN 的导游负面评论识别模型在验证集上的准确率可以达到 91.62%,损失为 0.25。在测试数据中的正面评论数据和负面评论数据上准确率分别为 0.916 4 和 0.919 7,召回率分别为 0.906 5 和 0.928 3, F1-score 分别为 0.911 4 和 0.923 9。

通过准确率、召回率和 F1-score 在相同数据集上对基于 TextCNN 导游负面评论识别方法、基于 CNN 的导游负面评论识别方法、基于 RNN 的导游负面评论识别方法、基于 LSTM 的导游负面评论识别方法和基于 FastText 的导游负面评论识别方法进行比较,如图 4 所示。综合比较 5 种方法,文中方法的准确率(0.916 2)、召回率(0.918 2)、F1-score(0.918 1)均高于其他方法,在验证集上的表现更加优秀。

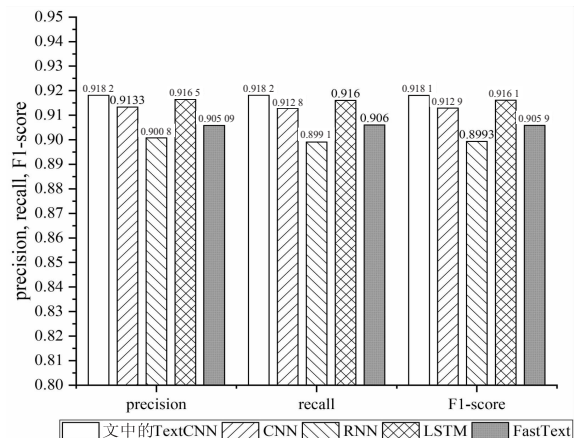


图 4 不同导游负面评论识别方法比较

#### 3.2.2 导游违规行为检测分类模型评估与分析

经过简单数据增强后共获得训练数据 9 000 余条,包括对强迫消费/参加自费项目行为的投诉 2 280 条,对擅自变更行程行为的投诉 1 910 条,对餐饮/住宿条件与合同不一致行为的投诉 1 570 条,对不具备从业资格行为的投诉 1 096 条,对殴打/辱骂游客行为的投诉 2 286 条。按照 0.8 的比例划分训练集和测试集,并在训练过程中在训练集中按照 0.8 的比例抽取数据进行验证。基于 LightGBM 的导游违规行为检测分类模型训练过程中,该文设定迭代 500 次。

基于 LightGBM 的导游违规行为检测分类模型在投诉信息的测试集上的正确率可以达到 88%,表 2 列出了该模型在测试集上各类别上的表现。使用准确率、召回率和 F1-score 在相同数据集上对基于 LightGBM 的导游违规行为检测分类方法、基于 XGBoost 的导游违规行为检测分类方法、基于逻辑斯特回归的导游违规行为检测分类方法和基于 SVM 的



导游违规行为检测分类方法进行比较,如图 5 所示。

表 2 基于 LightGBM 的导游违规行为  
检测模型各类别 F1-score

类别	F1-score
强迫消费/参加自费项目	0.924 7
擅自更改行程	0.820 5
餐饮/住宿条件与合同不一致	0.788 5
不具备相关从业资格	0.830 2
侮辱/殴打游客	0.916 7

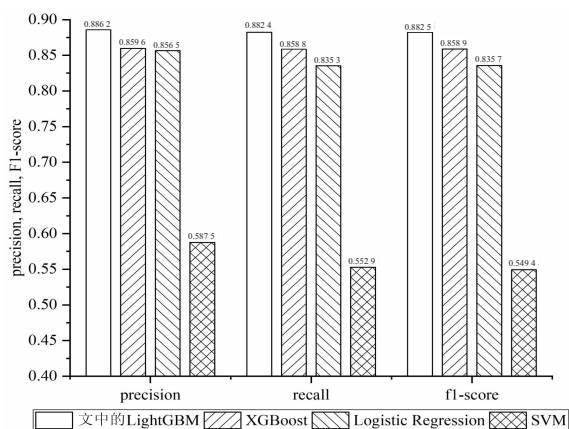


图 5 不同导游违规行为识别方法比较

由图 5 可知,文中方法的准确率最高,略高于 XGBoost 和 Logistic Regression,SVM 最差。

因为该文使用的数据集本身存在不平衡的特点,在参考 F1-score 外同时使用 ROC 曲线对不同方法进行评估,如图 6 所示。一个分类器的 ROC 曲线越靠近左上方,分类效果越好。

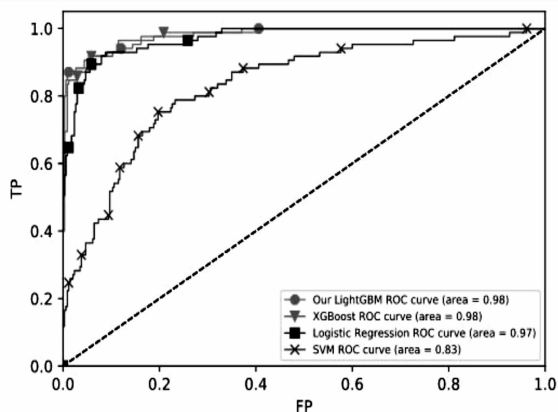


图 6 不同导游违规行为识别方法 ROC 曲线

由图 6 可知,LightGBM 和 XGBoost 的 AUC 面积最大,略高于 Logistic Regression,SVM 的 AUC 面积最小。考虑到 LightGBM 算法本身由 XGBoost 改进而来,在准确率相当的情况下训练时间更短(LightGBM 训练需要 59.12 s,XGBoost 训练需要 307.97 s)且内存占用更低,在文中的应用场景下 LightGBM 更加优秀。

由表 2 可知,由于“强迫消费”和“殴打/辱骂游

客”这两类数据中,带有强烈特征的描述词汇较多,模型识别效率较高。而在其他类别中,由于特征词汇较少,且样本数据存在一定的数据不均衡的问题,模型容易产生误判,F1-score 会稍低。

### 3.3 系统测试

文中方法应用到了实际的旅游大数据系统中,进行了相应的系统测试。使用的实际系统数据中包含导游相关的好评以及各类差评各 480 条。系统测试数据表明,文中方法可以达到 91.57% 的准确率(precision)、91.64% 的召回率(recall)、91.17% 的 F1-score 和 91.46% 的正确率(accuracy)。该系统对导游违规行为检测识别的示例如表 3 所示。文中方法适合于旅游大数据场景,能准确有效地识别检测出导游违规行为。

表 3 导游违规行为识别结果示例

类别	评论内容	违规行为
正常	…各景点内容详实。下次来成都报团游的话,希望还能碰到这样的导游…	/
	…适合喜欢慢生活的我们。小都导游前一天晚上及时联系游客,第二天准时接到我们…	/
	…今天跟的是小茜的团,全程都很舒服,时间把握的也好,不会很累,很舒服…	/
异常	…回到车上左导占据麦克风优势为自己辩解,并攻击游客,这是一种什么行为?	辱骂殴打游客
	…导游陈贵出言不逊、强制消费;故意拖延购物和停车时间…	强迫消费
	…导游证不给我们看…	不具备导游资质
	…增加自费路线,导致计划内行程严重缩水,并导致返程严重延期。该导游增加了 2 条计划外路线…	擅自修改行程

## 4 结束语

采用文本分类解决了从旅游评论文本中识别出检测导游违规行为的问题,构建了基于 TextCNN 的导游负面评论识别模型,进行了旅游评论的情感分析,识别出负面评论,再送入基于 LightGBM 的导游违规行为检测分类模型,实现对导游违规行为的检测与分类。通过和其他的主流模型的对比,该模型具备更好的性能,能够兼顾识别检测的准确率和召回率,降低正常行为被误判的可能性,适合于对旅游评论文本的导游违规行为进行检测。实验表明,提出的基于 TextCNN 和 LightGBM 的导游违规行为检测方法应用在实际旅游大数据系统中可以得到 91.57% 的准确率。构建的系统具备较高的可扩展性。在从旅游评论中检测出负面

评论后,对导游违规行为分类进行了初步探索。下一步还可以探索其他方面的投诉信息,进一步完善面向旅游市场的大数据智慧监管体系。

#### 参考文献:

- [1] 冯文雅. 2018 年全国实现旅游总收入 5.97 万亿元 同比增长 10.5% [EB/OL]. (2019-02-13). [http://www.xinhuanet.com/local/2019-02/13/c\\_1210058734.htm](http://www.xinhuanet.com/local/2019-02/13/c_1210058734.htm).
- [2] 仲召红. 新时期导游人员职业素质提升路径研究[J]. 池州学院学报, 2018, 32(3): 96-98.
- [3] KOWSARI K, MEIMANDI J, HEIDARYSAFA M, et al. Text classification algorithms: a survey [J]. Information, 2019, 10(4): 150.
- [4] ZHANG L, WANG S, LIU B. Deep learning for sentiment analysis: a survey [J]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2018, 8(4): e1253.
- [5] 梁昕露, 李美娟. 电信业投诉分类方法及其应用研究[J]. 中国管理科学, 2015, 23(Special Issue): 188-192.
- [6] 余本功, 陈杨楠, 杨颖. 基于 nBD-SVM 模型的投诉短文本分类[J]. 数据分析与知识发现, 2019, 3(5): 77-85.
- [7] 郑诚, 洪彤彤, 薛满意. 用于短文本分类的 BLSTM\_MLPCNN 模型[J]. 计算机科学, 2019, 46(6): 206-211.
- [8] 万圣贤, 兰艳艳, 郭嘉丰, 等. 用于文本分类的局部化双向长短时记忆[J]. 中文信息学报, 2017, 31(3): 62-68.
- [9] 段立, 徐鸿宇, 王懿, 等. 基于 word2vec 和 XGBoost 相结合的国网 95598 客服投诉工单分类[J]. 电力大数据, 2019, 22(12): 50-57.
- [10] 黄贤英, 刘广峰, 刘小洋, 等. 基于 word2vec 和双向 LSTM 的情感分类深度模型[J]. 计算机应用研究, 2019, 36(12): 3583-3587.
- [11] KIM Y. Convolutional neural networks for sentence classification [C]//Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). Doha, Qatar: [s. n.], 2014: 1746-1751.
- [12] ZHANG Y, WALLACE B C. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification [C]//Proceedings of the eighth international joint conference on natural language processing (volume 1: long papers). Taipei: [s. n.], 2017: 253-263.
- [13] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality [C]//Advances in neural information processing systems. Lake Tahoe, Nevada, USA: [s. n.], 2013: 3111-3119.
- [14] RONG X. Word2vec parameter learning explained [J]. arXiv: 1411.2738, 2014.
- [15] 李艳霞, 柴毅, 胡友强, 等. 不平衡数据分类方法综述 [J]. 控制与决策, 2019, 34(4): 673-688.
- [16] KE G, MENG Q, FINLEY T, et al. Lightgbm: a highly efficient gradient boosting decision tree [C]//Advances in neural information processing systems. Long Beach, California, USA: [s. n.], 2017: 3146-3154.
- [17] WEI J, ZOU K. EDA: easy data augmentation techniques for boosting performance on text classification tasks [C]//Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP). Hong Kong, China: [s. n.], 2019: 6383-6389.
- [18] SUN X, LIU M, SIMA Z. A novel cryptocurrency price trend forecasting model based on LightGBM [J]. Finance Research Letters, 2020, 32: 101084.
- [19] SHI X, CHENG Y, XUE D. Classification algorithm of urban point cloud data based on LightGBM [J]. IOP Conference Series: Materials Science and Engineering, 2019, 631(5): 052041.