

## ◆ 计算科学 ◆

## 基于 MacBERT 和联合注意力增强网络的物业服务投诉分类方法\*

湛志宏<sup>1\*\*</sup>, 覃开贤<sup>2</sup>, 彭凌华<sup>1</sup>, 湛 斌<sup>3</sup>

(1. 广西壮族自治区住房和城乡建设信息中心, 广西南宁 530028; 2. 南宁师范大学计算机与信息工程学院, 广西南宁 530001; 3. 西交利物浦大学, 江苏苏州 215028)

**摘要:** 基于人工的物业投诉文件分类处理方法已经无法满足社会需求, 并且已有投诉相关的自动分类方法在物业投诉分类问题上的性能较不足。因此, 本研究提出一个基于 MacBERT 和联合注意力增强网络的物业服务投诉分类方法 JAE-BERT4Com。JAE-BERT4Com 使用基于近义词替换与合成少数过采样技术结合的样本增强策略解决类不平衡的问题, 以及基于 MacBERT 的分层注意力、Transformers 的多头注意力和关键词注意力等多重注意力联合增强的网络进行文本特征学习和分类。实验结果表明, JAE-BERT4Com 能够获得比现有模型更高的准确率、F1 分数和召回率, 比现有较先进模型的性能更优。

**关键词:** 物业投诉; 投诉分类; 文本分类; 注意力增强; 深度学习

中图分类号: TP391 文献标识码: A 文章编号: 1005-9164(2024)01-0110-09

DOI: 10.13656/j.cnki.gxkx.20240417.011

近年来, 随着城市化进程的加速和人们生活水平的提高, 房地产与物业服务行业迅速发展。然而, 由于行业的特殊性和复杂性, 物业服务中常常会出现各种问题和投诉, 这些问题给房地产企业、物业管理公司、居民以及政府监管机构带来了不便和困扰, 也成为了一个重要的社会问题。通过文本投诉反馈到物业投诉平台, 是居民暴露物业相关问题、表达不满和寻求解决方案的重要方式, 这也成为物业管理人员和政府监管机构快速有效地解决物业问题、提高物业管

理水平和服务质量的重要途径<sup>[1]</sup>。然而, 随着投诉数量的增加, 人工分类处理投诉文件已经无法满足需求<sup>[1,2]</sup>。因此, 利用大数据和人工智能技术收集物业投诉数据, 对大规模的物业投诉内容进行有效的自动分类和辅助处理, 可大大提高对投诉文本的处理效率, 并发现一些潜在的问题和趋势, 对物业公司和政府监管机构制定针对性决策措施、改进服务质量、监管机制和优化运营具有重要的参考价值。

近年来, 国内外学者对多个领域服务投诉的自动

收稿日期: 2023-12-07

修回日期: 2024-01-04

\* 国家自然科学基金项目(62366011)资助。

## 【第一作者简介】

湛志宏(1972—), 男, 高级工程师, 主要从事计算机领域技术发展、数字政府、数字城市和住建系统信息化研究, E-mail: 410423205@qq.com。

## 【\*\*通信作者】

## 【引用本文】

湛志宏, 覃开贤, 彭凌华, 等. 基于 MacBERT 和联合注意力增强网络的物业服务投诉分类方法[J]. 广西科学, 2024, 31(1): 110-118.

ZHAN Z H, QIN K X, PENG L H, et al. Classification Method of Property Service Complaints Based on MacBERT and Joint Attention Enhancement Networks [J]. Guangxi Sciences, 2024, 31(1): 110-118.

分类问题开展了一些研究。Ghazzawi 等<sup>[3]</sup>分别用朴素贝叶斯、K 近邻、决策树和随机森林等模型进行投诉文本分类,这些方法对美国大都会交通管理局的交通服务投诉数据分类均有良好的性能。段立等<sup>[4]</sup>采用 Word2vec 和 XGBoost 相结合的方法实现对国家电网 95598 客服投诉工单的精准分类。杨颖等<sup>[5]</sup>综合考虑客户投诉中的文本信息及客户通讯状态信息,基于 Random Subspace 和支持向量机,并采用证据推理规则为集成策略构造分类模型对电信客户投诉进行分类。赵进等<sup>[6]</sup>为了自动分类投诉文本信息并将其落实到具体责任部门,首先通过 GRW 模型对投诉文本进行特征选择,提取有效特征词,然后构建基于 FastText 模型的用户投诉文本分类方法。范青武等<sup>[7]</sup>采用果蝇优化算法优化参数的支持向量机建立文本分类模型,以提高环境污染投诉文本的分类精度。王金政等<sup>[8]</sup>基于多头注意力机制,融合 BERT (Bidirectional Encoder Representations from Transformers)、Text-CNN 和 BiLSTM 等深度神经网络模型,实现对电信客户投诉文本的分类;Hossain 等<sup>[9]</sup>使用逻辑回归和支持向量机等机器学习方法来分类和评估从美国消费者金融保护局网站收集的客户投诉数据,实验结果表明这些方法可以准确分类客户投诉。Wang 等<sup>[10]</sup>提出一种基于 BERT 模型的联合注意力增强网络用于公民投诉报告的文本分类,并在多个数据集上取得了较好的结果。

尽管如此,目前国内外对物业投诉文本自动分类和分析的相关研究较少。作为该领域较早的相关研究工作,汪旭<sup>[11]</sup>用卷积神经网络(Convolutional Neural Networks, CNN)构建建筑质量投诉文本分类模型,取得比朴素贝叶斯模型和支持向量机模型更好的性能。朱明等<sup>[12]</sup>提出了面向物业投诉的字符级短文本分类深度学习联合模型 CNBG。该模型通过将物业投诉工单文本转化为字符向量表示,并分别输入到 CNN 和双向门控循环单元 BiGRU 中提取特征,再将这些特征进行融合,最终实现物业投诉文本分类。Bazzan 等<sup>[13]</sup>提出一个基于机器学习和分层缺陷分类的推荐系统,为建筑业与物业客户投诉提供了一个有效的智能分类方案。

总的来说,现有各种领域投诉文本分类方法主要有两类:一类是通过基于统计自然语言处理模型,如术语频率-逆文档频率(TF-IDF)的词频统计方法<sup>[14]</sup>学习得到词向量,然后输入朴素贝叶斯、K 近邻、决策树、随机森林、支持向量机模型等传统机器学习分类

模型进行文本分类;另一类是利用深度学习模型的文本分词结果或不分词直接将文本以字符为单位进行文本特征学习和分类的一体化处理。前一类方法需要人工标注特征,因此常因汉语的一词多义现象多和表达边界模糊等问题难以精准地分词和构建词向量,而且这类方法的上下文语境捕捉能力不足,无法应对大规模的投诉文本和内容复杂的文本数据,此外,这类方法还存在高维度、高稀疏等问题。深度学习模型在解决这些问题上表现出更好的效果,是未来发展的方向。但是,现有的基于深度学习的投诉文本分类模型倾向于以词向量的形式挖掘整个句子的语义,从而忽略了对整个句子学习中不同词的关注。各领域现有的大部分投诉文本分类模型都把投诉文本分类问题视为多分类问题求解,实际上物业投诉等一些领域投诉的文本分类并非多分类问题而是属于多标签分类问题,即一个物业投诉案件可同时属于多个投诉主题分类。此外,物业投诉案件存在的分类不平衡、事件场景多、分类边界模糊(可以由多个部门处理)和依赖关键词表达信息等特点也使分类任务更具挑战性。

针对以上问题,本研究提出一种使用基于近义词替换与合成少数过采样技术结合进行样本增强,并基于 MacBERT<sup>[15]</sup>和联合注意力增强网络的物业投诉分类方法 JAE-BERT4Com。通过加强对句子和投诉案件全文学习中不同词的关注和全局语义特征的理解,JAE-BERT4Com 能较好地解决分类不平衡问题,并在实验中表现出较好的性能。

## 1 相关工作

当今,预训练深度学习模型已成为自然语言处理(NLP)领域中流行的技术。BERT 是 Google 基于 Transformers 网络提出的专用于自然语言处理的开源预训练模型<sup>[16]</sup>。它使用未标记的大型纯文本语料库进行预训练,并用一种掩蔽语言建模(MLM)的方法来保持单词的焦点,从而能够通过周围的文本来建立上下文语境模式,进而帮助计算机理解文本中复杂的语言含义。BERT 相对传统的基于统计语言学方法和词嵌入的文本学习模型具有更强的理解语言上下文语境和处理歧义的能力。

然而,将 BERT 直接迁移到中文场景时会面临中英文语言差异,性能会显著下降。为解决此问题,Cui 等<sup>[15]</sup>在 BERT 的基础上提出了 MacBERT 模型(图 1)。MacBERT 缓解了 BERT“预训练-下游任

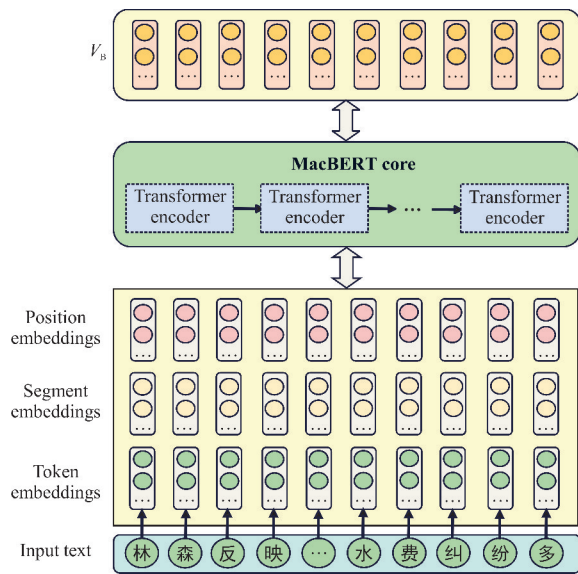


图1 MacBERT 框架

Fig. 1 MacBERT framework

务”不一致等问题,也更好地适配了中文场景<sup>[15]</sup>。MacBERT引入了一种纠错型掩码语言模型(Mac)预训练任务,使用了包括WWM(Whole Word Masking)、NM(N-gram Masking)、相似词替换等掩码方案,并在Tokenize模块中嵌入了中文分词功能,提升了预训练模型在多种NLP任务上的效果。由于MacBERT的主要框架与BERT完全一致,因此,本研究使用MacBERT进行物业投诉文本的分词处理

和向量化嵌入,并结合其他方法构建注意力增强网络。

## 2 基于MacBERT和联合注意力增强网络的物业投诉文本分类方法

### 2.1 JAE-BERT4Com的总体框架

JAE-BERT4Com的工作流程分为数据预处理、特征学习和投诉文本分类3个阶段(图2)。在数据预处理环节中,JAE-BERT4Com首先删除重复样本和主要内容不完整的无效样本,然后利用MacBERT内的Tokenize模块指导文本分词,最后利用近义词替换和合成少数过采样技术SMOTE<sup>[17]</sup>结合的方法进行样本增强。在特征学习环节中,JAE-BERT4Com首先用物业投诉语料库数据对MacBERT模型进行微调;然后将每句文本MacBERT的词嵌入送到关键词注意网络进行特征学习,同时将MacBERT输出的词向量输送到基于MacBERT的分层注意力网络进行特征学习。在投诉文本分类环节中,JAE-BERT4Com首先将两个注意力网络学习到的文本特征进行拼接,然后将拼接结果送入以Sigmoid交叉熵损失函数作为输出层的多层感知器,最终实现物业投诉的多标签分类。

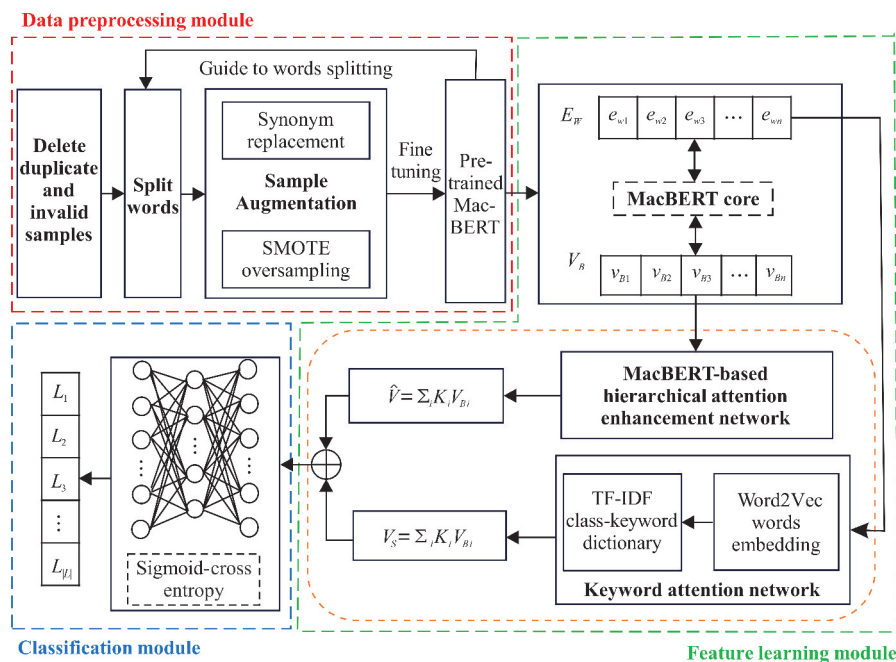


图2 JAE-BERT4Com的语义框架

Fig. 2 Semantic framework of JAE-BERT4Com

## 2.2 数据预处理

投诉信息样本从公开的线上投诉平台采集,由于用户操作不当等原因可能存在重复提交的投诉案件样本和基本内容不完整的投诉案件样本,故数据预处理需要首先删除这两类样本;然后需要删除特殊字符和无意义字符(如有些投诉者喜欢用连续多个“!”或“?”表达强烈的情绪,预处理时可保留一个符号);接着对文本进行分词处理。由于中文文本中只有句和段存在明显的分隔标志,不能像英文那样通过单词之间使用空格来实现分词,故 JAE-BERT4Com 算法在删除重复样本后调用 MacBERT 的 Tokenize 模块对投诉案件文本进行分词处理,并且用近义词替换和 SMOTE 过采样增强样本。最后,用 ChineseBERT 提供的 create\_pretraining\_data.py 将样本转换为 MacBERT 的样本编码<sup>[18]</sup>。

在机器学习领域,不平衡的训练数据集会导致模型的整体分类准确性较差。因此,针对物业投诉类不平衡的问题,本研究联合使用近义词替换与 SMOTE<sup>[17]</sup>来增强样本,以期解决类不平衡的问题。

### 2.2.1 近义词替换

汉语中有非常丰富的近义词,使用近义词替代中文语句中的少量词通常不会使其意思产生较显著的变化。因此,JAE-BERT4Com 算法将经数据预处理后得到的案件词表  $c$  (词的个数为  $n$ ) 加入样本库,同时从该案件词表中随机选择  $k$  个词替换成其相应的一个近义词,将替换后所得的案件词表  $\bar{c}$  作为一个新的物业投诉文本样本加入样本库,从而达到样本增强的效果。据本研究所用样本的统计分析结果,通常每个内容完整的物业投诉案件文本最少有 20 个文字,多则达 1 000 个文字。本研究假设替换少量 ( $1 \leq k \leq n \times 10\%$ ) 近义词不会导致该投诉样本的标签类别信息改变。近义词的查找使用中文近义词工具包 Synonyms 实现,形式化描述如式(1)所示。

$$c = \{c_1, c_2, c_3, \dots, c_i, \dots, c_n\} \rightarrow \bar{c} = \{c_1, c_2, c_3, \dots, \bar{c}_i, \dots, c_n\}, \quad (1)$$

其中,  $c_i$  为物业投诉案件词表  $c$  中的第  $i$  个词,  $\bar{c}_i$  为物业投诉案件词表  $c$  中的第  $i$  个词的一个近义词。

### 2.2.2 SMOTE 过采样

过采样是一种基于旧数据创建新的数据副本以平衡少数类中数据分布的技术,SMOTE 是一种通过复制少数类中的原始数据创建合成数据的过采样技

术<sup>[17,19]</sup>。SMOTE 通过寻找少数类中每个数据的最近邻来创建合成数据,旨在使少数类中的数据量等于多数类中的数据量,原理如式(2)所示<sup>[20]</sup>。

$$X_{\text{new}} = X_i + (\hat{X}_i - X_i) \times \delta, \quad (2)$$

其中,  $X_i$  表示少数类的特征向量,  $\hat{X}_i$  表示  $X_i$  最近邻,  $\delta$  表示 0 到 1 的数值。

与近义词替换进行样本增强不同, JAE-BERT4Com 中的 SMOTE 使用两个或多个投诉文本来创建新的数据。新合成的投诉文本采用新句型的形式呈现,没有使用在原始数据语料库之外的任何新词。在过采样过程中, JAE-BERT4Com 复制少数类中的所有样本数据合成新样本,直到与多数类中的数据样本量相同或尽可能相同。

## 2.3 物业投诉文本嵌入

为了让模型能理解输入的文本信息,需要先对输入的文本信息进行向量化嵌入。因此, JAE-BERT4Com 基于 MacBERT 和 Word2vec<sup>[21]</sup> 的文本嵌入以实现联合注意力学习。

### 2.3.1 基于 MacBERT 的文本嵌入

本节将经样本增强后所得的样本输入 MacBERT 预训练模型对其进行下游领域任务微调,让其更适合物业投诉文本主题分类。然后,用微调过的 MacBERT 模型对物业投诉文本样本进行学习并输出文本嵌入  $V_B = \{v_{B1}, v_{B2}, \dots, v_{Bn}\}$  和更新词嵌入  $E_W = \{e_{w1}, e_{w2}, \dots, e_{wm}\}$ 。

### 2.3.2 基于 Word2vec 的文本嵌入

Word2vec 是一个压缩词向量维度并结合词间关系的深度学习模型(图 3),其包含了 skip-gram 模型及连续词袋模型的思想<sup>[21,22]</sup>。设  $x, y, z$  分别是其上下文环境变量,  $W_{v \times n}$  和  $Q_{n \times v}$  是映射矩阵,  $V_c$  是查询词的嵌入向量,  $e$  是预测词的 one-hot 编码向量,则

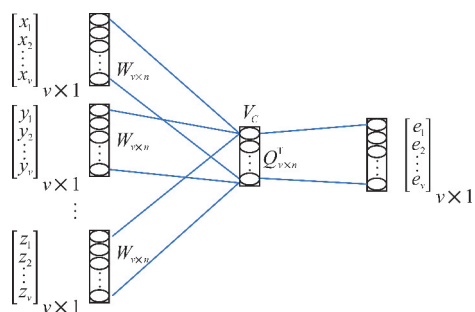


图 3 Word2vec<sup>[22]</sup>示意图

Fig. 3 Schema diagram of Word2vec

Word2vec 主要计算过程如下:

$$V_c = \frac{1}{3} \times W_{v \times n}^T (x + y + z), \quad (3)$$

$$e = \text{Softmax}(V_c Q_{v \times n}^T). \quad (4)$$

本研究利用 Word2vec 方法对 MacBERT 的词嵌入  $E_w = \{e_{w_1}, e_{w_2}, \dots, e_{w_m}\}$  做进一步嵌入处理, 以便构建类-关键词信息字典。

#### 2.4 构建类-关键词信息字典

经笔者的调研分析发现, 每个投诉的类别往往都有一些相关性较强的关键词。本研究利用此性质, 提出关键词注意力增强的策略, 即在对物业投诉文本的特征学习过程加强对关键词的注意力, 以期获得对投诉主题分类更有意义的文本特征。

本研究使用 TF-IDF 的统计方法获取关键词信息, 并构建类-关键词信息字典。TF-IDF 的核心思想为, 如果某个词在文档中出现的次数几乎与它在整个语料库中出现的次数一样多, 那么该词最有可能是该文档的关键词。因此, 一个单词的重要性与其在文档中出现的次数成正比, 与其在语料库中出现的频率成反比。术语频率 (TF) 表示单词在文档中出现的频率, 但为保证词频不会因文档中字数较多而产生较大值, 需要规范化处理, 第  $i$  个词  $w_i$  在文档  $d_j$  中的 TF 计算方法如下:

$$tf_{ij} = \frac{n_{ij}}{\sum_k n_{kj}}, \quad (5)$$

其中,  $n_{ij}$  表示词  $w_i$  在文档  $d_j$  中出现的次数,  $k = |d_j|$ 。

一个词的反文档频率 (IDF) 可以用文档总数除以包含该词的文档数来表示, IDF 越大说明该词的类别差异越好, 第  $i$  个词  $w_i$  在文档  $d_j$  中的 IDF 计算公式为

$$idf_i = \log \frac{|D|}{1 + |\{j: w_i \in d_j\}|}, \quad (6)$$

其中,  $|D|$  表示语料库中的文档总数,  $|\{j: w_i \in d_j\}|$  表示包含单词  $w_i$  的文档数, 为防止分母为零, 一般使用  $1 + |\{j: w_i \in d_j\}|$  作为分母。因此, 第  $i$  个词  $w_i$  的  $tf\_idf$  值为

$$tf\_idf_i = \frac{n_{ij}}{\sum_k n_{kj}} \cdot \log \frac{|D|}{1 + |\{j: w_i \in d_j\}|}. \quad (7)$$

本研究将物业投诉语料库中的样本归为  $m$  类, 并利用式(7)、(8)计算每个词的  $tf\_idf$  值并构建每个投诉主题类别  $L$  的类-关键词信息字典 Dis:

$$\text{Dis} = L_1 \cup L_i \cup \dots \cup L_m, \quad (8)$$

其中,  $L_i = \{(w_1: t_{1,i}), (w_2: t_{2,i}), (w_3: t_{3,i}), \dots, (w_n: t_{n,i})\}$  表示第  $i$  ( $1 < i \leq m$ ) 类的关键词信息字典,  $w_i$  表示语料库中的第  $i$  个关键词,  $t_{i,j}$  表示第  $i$  个关键词在第  $j$  类中的  $tf\_idf$  值。本研究将各个关键词在各类中的最大值作为其最终的  $tf\_idf$  值, 即

$$L[w_i] = \max(L_1[w_i], L_2[w_i], \dots, L_m[w_i]). \quad (9)$$

借此可将各个类别词典统一起来, 构建整个语料库的关键词信息字典。

#### 2.5 关键词注意力增强

根据上述的 Word2vec 的文本嵌入信息和类-关键词信息字典 Dis 信息, 用式(10)构建投诉文本关键词注意力增强后的句子增强向量  $V_s$ :

$$V_s = \text{Norm}(\text{TfIdf}(w_1))x^T + \text{Norm}(\text{TfIdf}(w_2))y^T + \dots + \text{Norm}(\text{TfIdf}(w_v))z^T, \quad (10)$$

其中,  $\text{TfIdf}(w_i)$  为词  $w_i$  在类-关键词信息字典 Dis 中的  $tf\_idf$  值;  $\text{Norm}(\cdot)$  为归一化操作;  $x \in R^{1 \times v}$ ,  $y \in R^{1 \times v}$ ,  $z \in R^{1 \times v}$  为词的嵌入向量;  $v$  为句输出向量的维数, 为了方便计算, 本研究设置 Word2vec 及 MacBERT 的输出向量维度也为  $v$ 。式(10)可视为通过类-关键词信息系数向词嵌入向量添加软注意力, 注意力系数为  $\text{Norm}(\text{TfIdf}(w_i))$ , 生成的句子增强向量  $V_s$  会更倾向于句子中的关键词。

#### 2.6 基于 MacBERT 的分层注意力

同一个词在不同的语境下所表达的意义会存在较大的差异, 这可能会对分类产生较大影响。同一句话中不同词构成的不同语境通常容易被 MacBERT 模型的多头注意力机制捕捉, 然而, 跨句的语境就难以被其捕捉。因此, 本研究提出句子注意力增强学习以捕捉跨句的语境对词分类特征的影响信息, 并奖励能够正确对案件进行分类的句子, 即在每个句子的 MacBERT 模型的输出向量  $V_B$  中使用双向 GRU 编码器对其中的每个词分量重新编码表达句子, 编码方式见式(11)<sup>[23]</sup>:

$$\begin{aligned} \vec{h}_i &= \overrightarrow{\text{GRU}}(V_{B_i}), i \in \{1, 2, \dots, |L|\}, \\ \overleftarrow{h}_i &= \overleftarrow{\text{GRU}}(V_{B_i}), i \in \{|L|, |L|-1, \dots, 1\}, \end{aligned} \quad (11)$$

其中,  $V_{B_i}$  表示 MacBERT 模型输出的案件文本  $V_B$  第  $i$  个句子向量,  $\vec{h}_i$  和  $\overleftarrow{h}_i$  分别表示 Bi-GRU 前、后向学习到的隐特征向量。将二者拼接得到第  $i$  个句子的总体隐特征向量  $h_i = \vec{h}_i \oplus \overleftarrow{h}_i$ ,  $h_i$  不仅包含第  $i$  个

句子信息特征,还包含了邻近句子的特征信息。然后,句子注意力系数  $a = (a_1, a_2, \dots, a_n)$  的计算方法如下:

$$a_i = \frac{\exp(uT_i u_s)}{\sum_i \exp(uT_i u_s)}, \quad (12)$$

其中,  $a_i$  为注意力系数的第  $i$  个分量,  $u_i = \tanh(\omega_B v_{B_i} + b_B)$ ,  $\omega_B$  和  $b_B$  分别表示权重和偏置,  $u_s$  表示句子级上下文向量,用来衡量句子的重要性。 $u_s$  可以随机初始化,并在模型训练过程中共同学习。综上,投诉案件文本向量  $\hat{V}$  可通过式(13)获得:

$$\hat{V} = \sum_i a_i h_i. \quad (13)$$

## 2.7 基于多重注意力联合增强的物业投诉文本主题分类

虽然 MacBERT 模型的句输出向量  $V_B$  的生成经过了 Transformersblock 中的多头注意力机制计算,但是 MacBERT 在投诉文本特征学习过程中,对具有较突出的分类代表性的类关键词关注不足,且对词在语句间的关联的挖掘较有限。为此,本研究提出一个基于多重注意力联合增强特征学习的策略,将前文中的结果拼接起来获得投诉案件文本的最终特征  $V = \hat{V} \oplus V_s$ ,并将其输送到多层感知器中进行分类预测(图 1 橙色虚线框内部分)。

由于多标签分类问题是一个类别非互斥的多分类问题,因此不能像二分类和多分类问题用类别互斥的 Softmax 函数作为激活函数。Sigmoid 函数的输出概率取值在 0 到 1 之间,但其和不一定等于 1(如一个物业投诉可以分类为“秩序管理”,也可以分类为“安全投诉”,二者可以同时存在)。因此,将其与交叉熵损失函数结合可解决物业投诉的多标签分类问题。

综上,对于共有  $|L|$  种主题类别的投诉分类问题,可在模型最后的全连接层定义  $|L|$  个输出神经元,并使用 Sigmoid 函数作为激活函数。最后使用 Sigmoid 交叉熵损失作为模型训练的目标损失函数,如式(14)所示:

$$\text{Loss} = \sum_{i=1}^{|L|} p(x_i) \log\left(\frac{\exp(x_i)}{1 + \exp(x_i)}\right), \quad (14)$$

其中,  $p(x_i)$  是样本  $x_i$  的真实类别概率分布。

## 3 实验与结果分析

### 3.1 数据集

因当前物业投诉分类的相关研究极少,暂无公开

可用数据集。本研究所用的语料库(共有 38 240 个物业投诉案件样本)和实验用数据集(共有 7 599 个物业投诉案件样本)均来自某地住建管理部门提供的物业投诉案件数据。这些数据覆盖“行政管理”“秩序管理”“安全管理”“环境卫生”“安全投诉”“质量投诉”“维修投诉”和“服务态度和质量”等 8 个主题类型。将实验用数据集分为两部分,其中的 80% 作为训练数据,其余 20% 作为测试数据。

### 3.2 评价指标与基线模型

为了评估 JAE-BERT4Com 的性能,本研究用准确率(Accuracy)、F1 分数(F1 score)和召回率(Recall)作为评估指标,如式(15)–(17)所示:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}}, \quad (15)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad (16)$$

$$\text{F1 score} = 2 \times \text{Recall} \times \frac{\frac{\text{TP}}{\text{TP} + \text{FP}}}{\frac{\text{TP}}{\text{TP} + \text{FP}} + \text{Recall}}, \quad (17)$$

其中,TP(True Positive)是指预测为正样本,且实际为正样本的数量;FP(False Positive)是指预测为正样本,且实际为负样本的数量;FN(False Negative)是指预测为负样本,且实际为正样本的数量;TN(True Negative)是指预测为负样本,且实际为负样本的数量。

本研究选取了一系列较先进的投诉分类相关方法进行对比实验,所比较的方法包括 3 个非深度学习模型 TF-IDF+RF<sup>[3]</sup>、Word2vec+XGBoost<sup>[4]</sup> 和 TF-IDF+SVM+Random Subspace<sup>[5]</sup>,以及 5 个深度学习模型 GRW+FastText<sup>[6]</sup>、BERT+Text-CNN+BiLSTM+multi-headed attention mechanism<sup>[8]</sup>、BERT+joint attention enhancement network<sup>[10]</sup>、CNBG<sup>[12]</sup> 和 Word Level CNN<sup>[24]</sup>。

### 3.3 实验结果

#### 3.3.1 JAE-BERT4Com 在各主题下的分类性能

表 1 给出了 JAE-BERT4Com 在数据集中各主题分类上的性能情况。JAE-BERT4Com 在各个主题分类中都表现良好。其中,模型在“维修投诉”主题类型上获得最好的性能,其相应的准确率、召回率以及 F1 分数分别是 0.852、0.893 和 0.868;在主题类型“秩序管理”上获得最差的性能,其相应的准确率、召回率以及 F1 分数分别为 0.719、0.758 和 0.731。通过观察分析这类案件样本发现,“维修投诉”类投诉

案件的文本里普遍存在与“修”字相关的关键词,而“秩序管理”类投诉案件的文本里无普遍存在的关键词,而且样本数最少,这可能是致使模型在“秩序管理”类上分类性能较差的主要原因。模型在“安全管理”和“安全投诉”两类投诉案件中的分类性能表现出了召回率较高而准确率相对低的现象,笔者分析这两类案件样本的标签预测结果发现准确率相对低是由于预测结果中假阳性的样本数较多。

表 1 JAE-BERT4Com 的各主题分类结果

Table 1 Classification results for each theme of JAE-BERT4Com

主题类型 Theme class	样本数 # Sample	准确率 Accuracy	F1 分数 F1 score	召回率 Recall
Administrative supervision	756	0.838	0.842	0.847
Order management	110	0.719	0.731	0.758
Security management	911	0.781	0.841	0.881
Sanitation	898	0.849	0.876	0.892
Security complaints	1 191	0.798	0.849	0.883
Quality complaints	2 919	0.849	0.869	0.878
Repair complaints	1 222	0.852	0.868	0.893
Service attitude and quality	2 338	0.834	0.845	0.857

表 2 对比实验结果

Table 2 Results of comparison experiments

	方法 Method	准确率 Accuracy	F1 分数 F1 score	召回率 Recall
Not DL	TF-IDF+RF <sup>[3]</sup>	0.778	0.744	0.738
	Word2vec+XGBoost <sup>[4]</sup>	0.781	0.807	0.811
	TF-IDF+SVM+Random Subspace <sup>[5]</sup>	0.805	0.824	0.841
DL	GRW+FastText <sup>[6]</sup>	0.820	0.828	0.835
	BERT+Text-CNN+BiLSTM+multi-headed attention mechanism <sup>[8]</sup>	0.821	0.830	0.834
	BERT+joint attention enhancement network <sup>[10]</sup>	0.818	0.831	0.829
	CNMG <sup>[12]</sup>	0.811	0.810	0.804
	Word Level CNN <sup>[24]</sup>	0.813	0.821	0.824
	JAE-BERT4Com(Our method)	<b>0.842</b>	<b>0.855</b>	<b>0.861</b>

Note: the best results are in bold.

### 3.3.3 消融实验

为了探究 JAE-BERT4Com 的各部分对模型性能的影响,对 JAE-BERT4Com 进行了消融实验(表 3)。-A 是 JAE-BERT4Com 去除了近义词替换的样本增强部分的方法,-B 是将 JAE-BERT4Com 去除了 SMOTE 样本增强部分的方法,-AB 是 JAE-BERT4Com 同时去除两种样本增强部分的方法,-C 是 JAE-BERT4Com 去除关键词注意力增强网络的方法,-D 是 JAE-BERT4Com 去除分层注意力增强网络的方法,-CD 是 JAE-BERT4Com 同时去

### 3.3.2 对比实验

本节将 JAE-BERT4Com 与较先进模型在所有物业投诉文本分类中比较总体性能。各模型在数据集的平均测试结果如表 2 所示。JAE-BERT4Com 对投诉主题进行分类时,在各个评价指标上的结果都较好。其准确率、F1 分数和召回率分别为 0.842、0.855 和 0.861,均优于其他方法。JAE-BERT4Com 比效果最好的对比方法(BERT+Text-CNN+BiLSTM+multi-headed attention mechanism)的准确率、F1 分数和召回率分别提高了 2.56%、3.01% 和 3.24%;比性能较差的 TF-IDF+RF 方法的准确率、F1 分数和召回率分别提高了 8.23%、14.92% 和 16.67%。召回率最高的 2 个方法分别是 JAE-BERT4Com 和 TF-IDF+SVM+Random Subspace 方法,二者都利用了 TF-IDF 信息,这说明利用 TF-IDF 信息有利于提高分类的召回率。基于深度学习(DL)的分类方法的性能均优于基于非深度学习的分类方法。

除两种注意力增强网络的方法。

从表 3 可知,当 JAE-BERT4Com 分别去除 4 个关键部分后,其性能都显著下降,这表明 JAE-BERT4Com 的每个部分对模型性能都起到正作用(表 3)。在 JAE-BERT4Com 去除样本增强策略后其各项性能指标均下降较多,这说明样本增强对模型的改进作用较大。在 JAE-BERT4Com 去除关键词注意力增强网络或同时去除两种注意力增强网络后,其各项性能指标值均明显下降,尤其是召回率下降较多,这表明本研究的关键字注意力增强策略对分类性

能产生了明显的作用,特别是对召回率的提高起到较显著的作用。

表 3 消融实验结果

Table 3 Results of ablation experiments

方法 Method	准确率 Accuracy	F1 分数 F1 score	召回率 Recall
JAE-BERT4Com	<b>0.842</b>	<b>0.855</b>	<b>0.861</b>
-A	0.828	0.845	0.856
-B	0.829	0.844	0.853
-AB	0.821	0.838	0.844
-C	0.832	0.834	0.831
-D	0.831	0.843	0.840
-CD	0.824	0.828	0.827

Note; the best results are in bold.

## 4 结论

为准确分类物业投诉文本,本研究提出一种基于 MacBERT 和联合注意力加强网络的物业服务投诉文本自动主题分类方法 JAE-BERT4Com。JAE-BERT4Com 把物业投诉分类问题视为多标签学习问题求解,使用了基于近义词替换与 SMOTE 过采样技术结合的样本增强策略,基于 MacBERT 的分层注意力、Transformers 的多头注意力、关键词注意力增强等的多重注意力联合增强特征学习。实验结果表明,JAE-BERT4Com 比现有最先进方法的性能更优,准确率、F1 分数和召回率分别可提高 2.56%—8.23%、3.01%—14.92% 和 3.24%—16.67%。

尽管 JAE-BERT4Com 性能表现较好,但其作为深度学习模型,比传统机器学习方法需要更多的计算资源和运行时间。未来,将进一步研究和改进模型的性能和效率,通过物业服务投诉文本的自动主题分类结果为解决物业问题提供科学指导。

## 参考文献

- [1] BAEK S, JUNG W, HAN S H. A critical review of text-based research in construction: data source, analysis method, and implications [J]. *Automation in Construction*, 2021, 132, 103915.
- [2] SALAMA D M, EL-GOHARY N M. Semantic text classification for supporting automated compliance checking in construction [J]. *Journal of Computing in Civil Engineering*, 2016, 30(1): 04014106.
- [3] GHAZZAWI A, ALHARBI B. Analysis of customer complaints data using data mining techniques [J]. *Procedia Computer Science*, 2019, 163: 62-69.
- [4] 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.
- [5] 杨颖, 王珺, 王刚. 基于改进的 Random Subspace 的客户投诉分类方法 [J]. *计算机工程与应用*, 2020, 56(13): 230-235.
- [6] 赵进, 杨小军. 基于 GRW 和 FastText 模型的电信用户投诉文本分类应用 [J]. *电信科学*, 2021, 37(6): 125-131.
- [7] 范青武, 陈光, 杨凯. 基于改进果蝇优化算法的投诉举报文本分类 [J]. *计算机工程与设计*, 2022, 43(4): 1103-1110.
- [8] 王金政, 杨颖, 余本功. 基于多头协同注意力机制的客户投诉文本分类模型 [J]. *数据分析与知识发现*, 2023, 7(1): 128-137.
- [9] HOSSAIN M S, RAHMAN M F, ALARIFI G. Prediction and analysis of customer complaints using machine learning techniques [J]. *International Journal of E-Business Research*, 2023, 19(1): 1-25.
- [10] WANG Y H, ZHOU Y H, MEI Y D. A joint attention enhancement network for text classification applied to citizen complaint reporting [J]. *Applied Intelligence*, 2023, 53(16): 19255-19265.
- [11] 汪旭. 建筑质量投诉文本分类与知识问答系统研究 [D]. 武汉: 华中科技大学, 2018.
- [12] 朱明, 陈一飞. 面向物业投诉的字符级短文本分类模型 [J]. *大众科技*, 2022, 24(4): 31-35.
- [13] BAZZAN J, ECHEVESTE M E, FORMOSO C T, et al. An information management model for addressing residents' complaints through artificial intelligence techniques [J]. *Buildings*, 2023, 13(3): 737.
- [14] KIM D, SEO D, CHO S, et al. Multi-co-training for document classification using various document representations: TF-IDF, LDA, and Doc2Vec [J]. *Informing Sciences*, 2019, 477: 15-29.
- [15] CUI Y M, CHE W X, LIU T, et al. Pre-training with whole word masking for Chinese BERT [J]. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021, 29: 3504-3514.
- [16] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding [C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: ACL, 2019: 4171-4186.
- [17] SÁEZ J A, LUENGO J, STEFANOWSKI J, et al. SMOTE-IPF: addressing the noisy and borderline examples problem in imbalanced classification by a resampling method with filtering [J]. *Information Sciences*, 2015, 291: 184-203.
- [18] ÖZDEMİR A, POLAT K, ALHUDHAIF A. Classification of imbalanced hyperspectral images using SMOTE-based deep learning methods [J]. *Expert Systems with Applications*, 2021, 178: 114986.
- [19] INDRAHIMAWAN M R, SANTOSA P I, ADJI T B.



- Handling data imbalance using text augmentation for classifying public complaints [C]//2023 International Conference on Computer, Control, Informatics and its Applications (IC3INA). Bandung: IEEE, 2023: 284-289.
- [20] LISJANA O A, KHODRA M L. Classifying complaint reports using RNN and handling imbalanced dataset [C]//2022 9th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE). Semarang: IEEE, 2022: 303-307.
- [21] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality [C]//Proceedings of the 26th International Conference on Neural Information Processing Systems. Lake Tahoe: ACM, 2013: 3111-3119.
- [22] ZHU K F, MA B S, HUANG T H, et al. Sequence generation network based on hierarchical attention for multi-charge prediction [J]. IEEE Access, 2020, 8: 109315-109324.
- [23] YANG Z C, YANG D Y, CHRIS D, et al. Hierarchical attention networks for document classification [C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego: ACL, 2016: 1480-1489.
- [24] NAIK P K, PRASHANTH T, CHANDRU S, et al. Consumer complaints classification using machine learning & deep learning [J]. International Research Journal on Advanced Science Hub, 2023, 5(5S): 116-122.

---

## Classification Method of Property Service Complaints Based on MacBERT and Joint Attention Enhancement Networks

ZHAN Zhihong<sup>1\* \*</sup> , QIN Kaixian<sup>2</sup>, PENG Linghua<sup>1</sup>, ZHAN Cheng<sup>3</sup>

(1. Information Center of Housing and Urban-rural Development of Guangxi Zhuang Autonomous Region, Nanning, Guangxi, 530028, China; 2. School of Computer and Information Engineering, Nanning Normal University, Nanning, Guangxi, 530001, China; 3. Xi'an Jiaotong-Liverpool University, Suzhou, Jiangsu, 215028, China)

**Abstract:** The manual-based classification method of property complaint documents has been unable to meet the needs of the society, and the existing automatic classification methods related to complaints have insufficient performance in the classification of property complaints. Therefore, this study proposes a property service complaint classification method JAE-BERT4Com based on MacBERT and joint attention enhancement network. JAE-BERT4Com uses a sample enhancement strategy based on the combination of synonym replacement and synthetic minority oversampling technology to solve the problem of class imbalance. And a multi-attention joint enhancement network based on MacBERT's hierarchical attention, Transformers' multi-head attention and keyword attention is designed to perform text feature learning and classification. The experimental results show that JAE-BERT4Com can obtain higher accuracy, F1 score and recall rate than the existing models, and has better performance than the existing advanced models.

**Key words:** property complaints; complaint classification; text categorization; attention enhancement; deep learning

责任编辑: 陆 雁, 陈少凡

---