人工智能

基于预训练模型的为企服务专题事件识别

代晓菊
（上海理想信息产业（集团）有限公司，上海 201315）

摘　要：基于预训练模型的为企服务专题事件识别旨在通过技术手段和专业运营，提升12345热线平台的数据分析能力和常态化数据分析服务。结合12345市民热线在城市数字化治理中的新定位，利用BERT预训练模型在二分类和多标签分类任务中的应用，通过无监督学习从大规模文本语料中学习到丰富的语义表示，提升工单分类的准确率。并优化为企服务专题的构建过程，通过筛选企业相关数据，建立专题数据识别模型，对企业类型的工单进行事件抽取和结果分析，为政府决策提供了有力支持。

关键词：文本分类；预训练模型；自然语言处理
中图分类号：TP393
文献标志码：A
doi: 10.11959/j.issn.1000−0801.2025064

0　引言

12345市民热线自2012年开通运行以来，主体业务是受理市民来电，完成市民诉求的受理、转派和回访。与市民进行沟通互动的主要渠道是传统语音，在服务渠道方面，提供了电话语音和传真的方式受理市民诉求。围绕着为市民诉求解答、受理、反馈的闭环业务，热线中心形成了一套成熟的业务支撑平台。

近年来，在上海超大城市数字化治理的背景下，全市推进“一网统管”数字城运治理和“一网通办”数字政务管理的“两张网”建设，12345市民服务热线也获得了“三网总客服”新的定位，对热线的数智化升级、赋能城市治理等方面提出新的需求。

为进一步向政府决策提供数据支持，将真实诉求转化为建设驱动、短板弥补的依据，发掘数据价值，把钱花在刀刃上，让政府办事更得民心。本论文将围绕12345热线平台的数据分析能力和常态化数据分析服务，通过技术手段和专业运营，为热线办提供数据专题报告、事件发现等数字化服务能力。

1　文本分类算法

BERT是2018年10月由Google AI研究院提出的一种预训练模型。BERT的全称是bidirectional encoder representations from transformers。在SQuAD1.1中表现出惊人的成绩，两个衡量指标上全面超越人类，成为NLP发展史上的里程碑式的模型成就。

BERT是一种基于Transformer架构的预训练语言模型，通过无监督学习从大规模文本语料中学习到丰富的语义表示。BERT能够在上下文中理解词语的含义，从而对序列标注任务非常有用。

收稿日期：2024−11−15；**修回日期：**2025−02−05

1.1　二分类模型

基于 BERT 的二分类模型是一种使用 BERT 预训练模型作为特征提取器的分类算法。通过在大型文本语料上进行无监督的训练来学习丰富的词汇和语义信息，BERT 是一种强大的语言表达模式。以下是本项目为了识别企业类型的工单，使用 BERT 进行二分类任务的实现步骤。

步骤 1　数据准备。准备用于训练和评估的标记文本数据集。每个样本应包含一个文本输入和对应的二分类标签。

步骤 2　BERT 模型加载。从预训练的 BERT 模型中增加相应的配置文件、权重和词汇表。

步骤 3　输入编码。将文字顺序转换为可理解的输入格式。BERT 接受输入序列的标记化表示和每个标记的分段 ID 两个输入。对于二类分类任务，一般是以 cls 符号来表示整句话，将所有的文字都视为句子。

步骤 4　特征提取。向 BERT 模型反馈编码后的输入，以获得文本的隐层表示。BERT 模型会产生向量表示，向量表示可以被认为是截取文本语义的固定长度。

步骤 5　分类层。在 BERT 的输出上添加一个分类层，可以是全连接层或其他适合任务的分类器。该分类器将根据任务目标将隐层向量映射到二分类的预测概率。

步骤 6　训练。使用标注的训练数据对模型进行训练，通过计算损失函数（如交叉熵）来调整模型参数。可以使用梯度下降等优化算法来最小化损失。

步骤 7　评估和推断。利用评价数据对训练后的模型进行评价，并对精确度、召回率等指标进行计算。对于新的未标记文本，可以使用已训练的模型进行推断并预测其类别。

BERT 模型的强大之处在于它能够从大规模的无监督数据中学习语义信息，并且可以通过微调适应各种下游自然语言处理任务，包括二分类任务。在本项目开发过程中，发现 Bert 预训练模型能够提升工单分类的准确率，因此在本项目过程中仅使用 Bert 算法进行了为企服务中的工单类型识别的分类实验。

1.2　多标签分类算法模型

多标签分类的输出空间会随着标签的数量指标增长，为了应对指数复杂度的标签空间，标签之间的关联性需要挖掘。多标学习能否成功，关键在于有效挖掘标签之间的关联性。根据 Multi-Tag 分类算法采用的标签相关性，分为一级策略、二级策略、高级策略等分类算法。在为企服务中考虑多个标签之间的关联，比如考虑所有其他标签对每个标签的作用。显然，在计算方面要求较高的高阶策略，高阶策略的相关性建模能力要强于一二阶策略。

多标签分类算法（multi-tag classification）如果按照算法设计思路的来源，可以分为问题转化方式和算法改编方式两大类。基于题目转换的多标签分类法，一般会将多标签分类题转换成其他学习场景，如二分类题、标签排序题、多分类题（multi-lock）等。基于算法改编的多标签分类方法一般是通过对流行学习算法的改编，如对决策树的改编、对量机的支持等直接处理多标签数据，随着预训练模型的发展，本项目将把预训练模型与 CNN、RNN 等深度学习算法相结合，使之适用于多标签分类（multi-tag）。

如果根据算法设计思想的来源，可以将多标签分类算法分为两类：问题转换的方法和算法改编的方法。基于问题转换的多标签分类方法一般将多标签分类问题转换为其他学习场景，比如转换为二分类问题、标签排序问题、多分类问题等。基于算法改编的多标签分类方法一般是通过改编流行的学习算法去直接处理多标签数据，比如改编决策树、支持向量机等等，随着近些年预训练模型的发展，本项目将结合预训练模型与 CNN、RNN 等深度学习算法，使其适用于多标

签的分类。本项目中涉及的多标签分类模型的实现具体流程如下。

- 利用keras-ber加载预训练好的bert，这里用的bert是哈工大训练的chinese_bert_wwm_L-12_H-768_A-12。
- 取出bert的输出中的[cls]向量，[cls]可以直接用于分类，也可以与其他网络的输出拼接。
- 取出bert输出中关于输入句子的表示（word_embedding），bert在输入时在句子的头和尾分别添加了一个[CLS]、[SEP]，可以选择去除这两个标志。
- 将word_embedding输入构造好的多kernel size的TextCNN网络，获得经由TextCNN获得特征（cnn_features）。
- 将[cls]与cnn_features进行拼接后用于分类。
- 根据输入和输出封装模型，并进行必要参数的配置。

2 为企服务专题的构建过程

2.1 业务需求

在热线数字化转型升级的大背景下，针对当前热线现有系统缺乏专业的、规范化的数据分析能力。12345市民服务热线需要持续提高数据分析能力。在数字化转型的背景下，12345市民服务热线亟须通过AI数据分析能力，获得大量市民诉求的快速处理能力和问题挖掘分析能力。通过本项目的建设，可以基于AI模型识别工单内容中业务分类、事项标签以及关键热词等内容，获得按照专题、业务分类的聚类分析能力，让12345热线实现多维度的提质增效。

上海12345市民服务热线每天工单受理量约30 000单，每天工单派发量约8 000单。12345热线业务类型种类多样，且横跨多个部门及单位，如交通委、公安局、社保局、市卫生健康委等，还需要纵向深入市、区、街道、镇、居委等各级单位，工单流转复杂，基层工单受理人员压力巨大，通过对热线数据的业务分类识别、事项识别等，形成专属的专题AI模型，实现对热点事件、热点诉求、难点问题、堵点问题等内容的挖掘。

2.2 为企服务专题分析解决方案

2.2.1 整体流程

首先判断是否为企业相关工单，需要针对该类场景建立专题数据识别模型，判断企业类型的工单数据。具体包括，针对前期为企服务工单事件的样例积累，1～3月共积累为企服务事件标签数据48 000条，针对未打标工单进行人工补充，将该数据作为训练数据，基于bert预训练模型进行训练。然后，对定期待分析的企业数据进行分析，预测每条工单是否为企业，针对是企业的工单数据进行事件抽取，最终进行结果分析。整体流程如图1所示。

2.2.2 整体流程详解

为企服务流程如图2所示，具体步骤如下。

步骤1 对数据进行清洗和分类。对为企服务数据进行数据清洗，对一些未打标数据、标签不匹配数据进行清理，包括删除未打标空白数据、修改标签不统一数据等。再针对这些已经清洗完的数据打乱分类，其中80%的数据为训练集，20%的数据为测试集。

步骤2 使用训练集训练模型。通过设置不同的参数，如learning_rate（学习率）、batch_size（每个批量的大小）、epochs（迭代次数）等，可以得到多个训练模型。

步骤3 使用不同参数的模型在验证集上进行验证，选择验证效果最好的模型。

步骤4 对模型数据进行分析。在验证集上，对每类标签进行数据分析，将准确率偏低的标签数据取出，分析原因。如果是因为数据量少而出现准确率偏低的情况，则需要补充数据。如果本

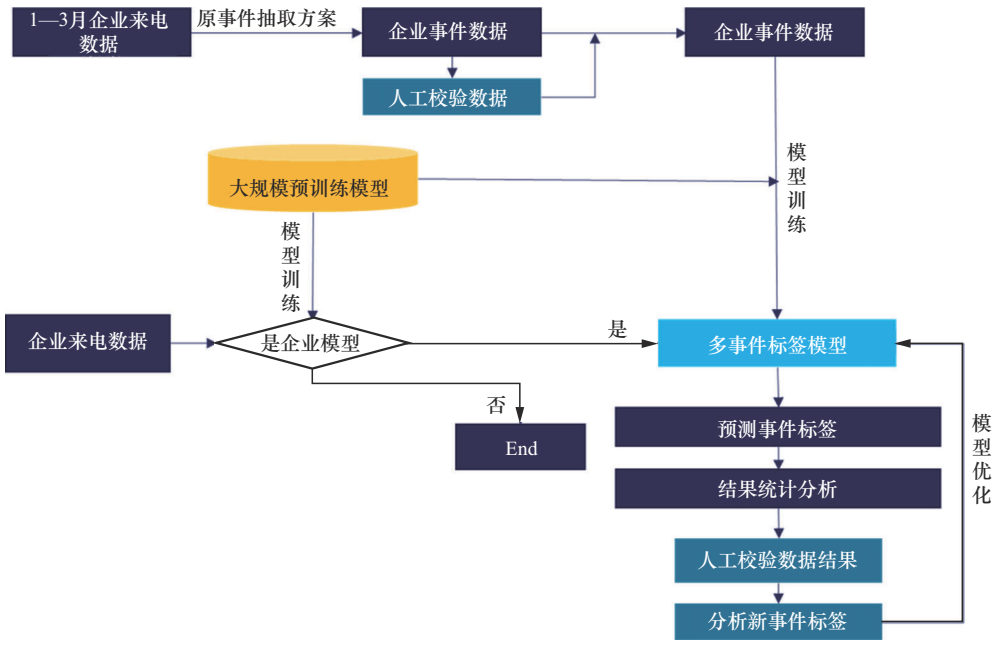

图1　整体流程

身准确率偏低，则需要人工打标团队进行打标确认，确认模型打标是否准确。

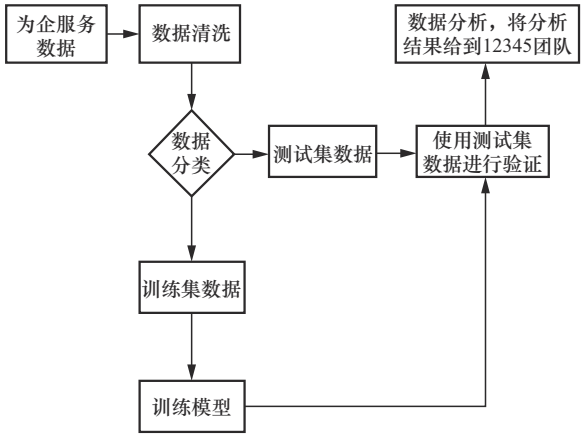

图2　为企服务流程

3　为企服务专题的数据结果分析

本功能上线后，测试集上准确率逐渐提高，其准确率核验标准，即模型预测的标签必须和人工所打标签一致才算准确。但实际上，模型预测的标签会比人工标签更加准确，包括人工打错标签、人工打漏标签等。虽然经过代码测算其准确率没有突破 85%，但实际上针对大数据量的标签准确率均突破 90%。同时，经过数据分析，发现数据中依然存在大量脏数据，需要时间对其进行清洗，因此准确率还会提高。相似度阈值与工单覆盖情况见表 1。

表 1　相似度阈值与工单覆盖情况

日期	数据量	准确率
8.09	3.2w	0.84
9.11	5.4w	0.85
10.10	6.1w	0.87

4　结束语

本次数据分析专题围绕为企服务的专题进行分析，建立为企服务事件标签体系，并建立热点事件标签模型，通过对工单数据进行专题识别、事项分类、热点分析等，对上海 12345 市民服务热线从电话、多媒体等渠道获取的市诉求信息进行全面数据分析，充分发挥 12345 市民服务热线在传递民情民意、为企惠企服务等方面的提升作用，助力政府决策。通过联合专题识别的二分类

模型以及事项分类的多标签识别模型，证实了这种方式可以用于进行专题识别的处理过程，大大提升了数据分析人员的服务效能。

参考文献：

[1] 李可悦, 陈轶, 牛少彰. 基于BERT的社交电商文本分类算法[J]. 计算机科学, 2021.

[2] 胡康. 基于BERT的热线平台案件自动分类研究与实现[D]. 湖南：湖南大学, 2022.

[3] 尹佳男. 政务热线如何践行"人民城市"理念[J]. 党政论坛, 2022.

[4] 崔雨萌,王靖亚,刘晓文等. 融合注意力和裁剪机制的通用文本分类模型[J]. 计算机应用, 2023.

[5] 汤雪. 基于深度学习的文本情感分类研究[D]. 广东: 华南理工大学, 2018.

[作者简介]

代晓菊（1990− ），女，上海理想信息产业（集团）有限公司工程师，主要从事NLP自然语言处理领域的研究和技术开发等工作。