(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 116089605 A
(43)申请公布日 2023.05.09

(21)申请号 202211490263.6

(22)申请日 2022.11.25

(71)申请人 海南大学
地址 570100 海南省海口市人民大道58号

(72)发明人 黄梦醒 潘志强 张文生 毋媛媛 冯思玲 冯文龙 张雨

(74)专利代理机构 苏州中合知识产权代理事务所(普通合伙) 32266
专利代理师 景晓玲

(51)Int.Cl.
G06F 16/35(2019.01)
G06F 18/23213(2023.01)
G06F 16/335(2019.01)

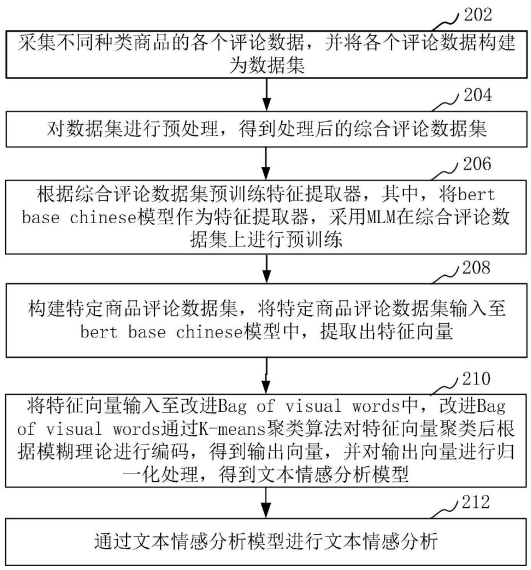
权利要求书2页 说明书11页 附图4页

(54)发明名称

基于迁移学习和改进词袋模型的文本情感分析方法

(57)摘要

本方案涉及一种基于迁移学习和改进词袋模型的文本情感分析方法、系统、计算机设备及存储介质。所述方法包括:采集不同种类商品的各个评论数据进行预处理后得到综合评论数据集;采用MLM根据综合评论数据集预训练特征提取器bertbasechinese模型;构建特定商品评论数据集,将特定商品评论数据集输入至bertbasechinese模型中提取出特征向量;将特征向量输入至改进Bagof visualwords中,通过K-means聚类算法对特征向量聚类后根据模糊理论进行编码,得到输出向量,并对输出向量进行归一化处理,得到文本情感分析模型;通过文本情感分析模型进行文本情感分析。通过迁移学习和Bagofvisualwords方法,能够很好的处理不断涌现的新类别商品的评论,降低文本情感分析的成本。



CN 116089605 A

1.一种基于迁移学习和改进词袋模型的文本情感分析方法,其特征在于,所述方法包括:

采集不同种类商品的各个评论数据,并将各个所述评论数据构建为数据集;

对所述数据集进行预处理,得到处理后的综合评论数据集;

根据所述综合评论数据集预训练特征提取器,其中,将bert base chinese模型作为特征提取器,采用MLM在所述综合评论数据集上进行预训练;

构建特定商品评论数据集,将所述特定商品评论数据集输入至所述bert base chinese模型中,提取出特征向量;

将所述特征向量输入至改进Bag ofvisual words中,所述改进Bag ofvisual words通过K-means聚类算法对所述特征向量聚类后根据模糊理论进行编码,得到输出向量,并对所述输出向量进行归一化处理,得到文本情感分析模型;

通过所述文本情感分析模型进行文本情感分析。

2.根据权利要求1所述的基于迁移学习和改进词袋模型的文本情感分析方法,其特征在于,所述将各个所述评论数据构建为数据集,包括:

将各个所述评论数据以csv的形式保存,且每条数据包含有类别、正负标签、评论。

3.根据权利要求1所述的基于迁移学习和改进词袋模型的文本情感分析方法,其特征在于,所述对所述数据集进行预处理,得到处理后的综合评论数据集,包括:

从所述数据集中取出各个所述评论数据的评论部分;

使用正则表达式的方式去掉各个所述评论部分中的无意义符号以及非中文内容,得到综合评论数据集。

4.根据权利要求1所述的基于迁移学习和改进词袋模型的文本情感分析方法,其特征在于,所述将所述特定商品评论数据集输入至所述bert base chinese模型中,提取出特征向量,包括:

通过Tokenizer工具对输入的所述特定商品评论数据集中的数据进行分词,并在分词后的样本上加上Token;

获取所述bert base chinese模型在预训练时的字典,并根据所述字典将各个所述Token映射为对应的ID;

通过所述bertbase chinese模型将所述特定商品评论数据集中映射为ID的等长样本转化为数值矩阵,并提取所述特定商品评论数据集中句子的语义特征和Token的上下文信息,经过输出层输出。

5.根据权利要求4所述的基于迁移学习和改进词袋模型的文本情感分析方法,其特征在于,所述采用MLM在所述综合评论数据集上进行预训练,包括:

从加上所述Token的样本中选取目标占比的目标Token;

选取第一阈值数量的所述目标Token替换为mask,选取第二阈值数量的所述目标Token替换为随机Token,选取第三阈值数量的所述目标Token保留。

6.根据权利要求1所述的基于迁移学习和改进词袋模型的文本情感分析方法,其特征在于,所述构建特定商品评论数据集,包括:

获取特定商品的评论数据并构建初步特定商品评论数据集;

对所述初步特定商品评论数据集中的评论数据用正则表达式的方式进行预处理,得到

处理后的特定商品评论数据集；

对所述特定商品评论数据集进行划分，构建训练集、验证集、测试集。

7.根据权利要求6所述的基于迁移学习和改进词袋模型的文本情感分析方法，其特征在于，所述改进Bag ofvisual words通过K-means聚类算法对所述特征向量聚类后根据模糊理论进行编码，得到输出向量，并对所述输出向量进行归一化处理，得到文本情感分析模型，包括：

从所述训练集中抽取训练样本，通过所述bert base chinese模型进行语义特征提取，对提取出的特征使用K-means聚类的方法，得到聚类中心列表；

对提取出的特征使用所述改进Bag ofvisual words进行编码，每个样本被编码为数值向量；

将所述特征向量转化为概率值。

8.一种基于迁移学习和改进词袋模型的文本情感分析系统，其特征在于，所述系统包括：

数据采集模块，用于采集不同种类商品的各个评论数据，并将各个所述评论数据构建为数据集；

预处理模块，用于对所述数据集进行预处理，得到处理后的综合评论数据集；

预训练模块，用于根据所述综合评论数据集预训练特征提取器，其中，将bertbase chinese模型作为特征提取器，采用MLM在所述综合评论数据集上进行预训练；

特征提取模块，用于构建特定商品评论数据集，将所述特定商品评论数据集输入至所述bert base chinese模型中，提取出特征向量；

模型训练模块，用于将所述特征向量输入至改进Bag ofvisual words中，所述改进Bag ofvisual words通过K-means聚类算法对所述特征向量聚类后根据模糊理论进行编码，得到输出向量，并对所述输出向量进行归一化处理，得到文本情感分析模型；

情感分析模块，用于通过所述文本情感分析模型进行文本情感分析。

9.一种计算机设备，包括存储器和处理器，所述存储器存储有计算机程序，其特征在于，所述处理器执行所述计算机程序时实现权利要求1至7中任一项所述方法的步骤。

10.一种计算机可读存储介质，其上存储有计算机程序，其特征在于，所述计算机程序被处理器执行时实现权利要求1至7中任一项所述的方法的步骤。

基于迁移学习和改进词袋模型的文本情感分析方法

技术领域

[0001] 本发明涉及自然语言处理技术领域,特别是涉及一种基于迁移学习和改进词袋模型的文本情感分析方法、系统、计算机设备及存储介质。

背景技术

[0002] 随着互联网技术的发展,网络购物群体数量在逐渐上升,截至2022年6月,网络购物用户规模达8.41亿,占网民整体的80%。面对如此之大的网络购物群体,在电商平台上每天都会产生数以亿计的评论,这些评论对于商家和消费者来说,都具有很高的参考价值。正确处理好这些宝贵的评论,对于营造一个良好的购物环境,具有重要作用。随着近年来自然语言理解领域突破性技术的涌现,学界对于文本情感分析的研究也越发的广泛而深入。传统的文本情感分析方式均是在一类商品的评论数据下训练好一个情感分析模型,从而得到该类商品评论的情感分析结果。

[0003] 然而,在一类商品的评论数据下训练好一个情感分析模型时,再想把这个模型应用到其他类型的商品,由于评论数据的分布存在差异,会导致模型的效果变差。同时由于深度模型需要训练的参数很多,所以要想重新训练一个模型,又需要花费极大的成本。因此,针对不同类型商品进行分析时需要训练对应的分析模型,存在成本较高的问题。

发明内容

[0004] 基于此,为了解决上述技术问题,提供一种基于迁移学习和改进词袋模型的文本情感分析方法,可以对不同类别商品的评论进行情感分析,降低了情感分析的成本。

[0005] 一种基于迁移学习和改进词袋模型的文本情感分析方法,所述方法包括:

[0006] 采集不同种类商品的各个评论数据,并将各个所述评论数据构建为数据集;

[0007] 对所述数据集进行预处理,得到处理后的综合评论数据集;

[0008] 根据所述综合评论数据集预训练特征提取器,其中,将bertbase chinese模型作为特征提取器,采用MLM在所述综合评论数据集上进行预训练;

[0009] 构建特定商品评论数据集,将所述特定商品评论数据集输入至所述bert base chinese模型中,提取出特征向量;

[0010] 将所述特征向量输入至改进Bag ofvisual words中,所述改进Bag ofvisual words通过K-means聚类算法对所述特征向量聚类后根据模糊理论进行编码,得到输出向量,并对所述输出向量进行归一化处理,得到文本情感分析模型;

[0011] 通过所述文本情感分析模型进行文本情感分析。

[0012] 在其中一个实施例中,所述将各个所述评论数据构建为数据集,包括:

[0013] 将各个所述评论数据以csv的形式保存,且每条数据包含有类别、正负标签、评论。

[0014] 在其中一个实施例中,所述对所述数据集进行预处理,得到处理后的综合评论数据集,包括:

[0015] 从所述数据集中取出各个所述评论数据的评论部分;

[0016] 使用正则表达式的方式去掉各个所述评论部分中的无意义符号以及非中文内容，得到综合评论数据集。

[0017] 在其中一个实施例中，所述将所述特定商品评论数据集输入至所述bert base chinese模型中，提取出特征向量，包括：

[0018] 通过Tokenizer工具对输入的所述特定商品评论数据集中的数据进行分词，并在分词后的样本上加上Token；

[0019] 获取所述bert base chinese模型在预训练时的字典，并根据所述字典将各个所述Token映射为对应的ID；

[0020] 通过所述bertbase chinese模型将所述特定商品评论数据集中映射为ID的等长样本转化为数值矩阵，并提取所述特定商品评论数据集中句子的语义特征和Token的上下文信息，经过输出层输出。

[0021] 在其中一个实施例中，所述采用MLM在所述综合评论数据集上进行预训练，包括：

[0022] 从加上所述Token的样本中选取目标占比的目标Token；

[0023] 选取第一阈值数量的所述目标Token替换为mask，选取第二阈值数量的所述目标Token替换为随机Token，选取第三阈值数量的所述目标Token保留。

[0024] 在其中一个实施例中，所述构建特定商品评论数据集，包括：

[0025] 获取特定商品的评论数据并构建初步特定商品评论数据集；

[0026] 对所述初步特定商品评论数据集中的评论数据用正则表达式的方式进行预处理，得到处理后的特定商品评论数据集；

[0027] 对所述特定商品评论数据集进行划分，构建训练集、验证集、测试集。

[0028] 在其中一个实施例中，所述改进Bag ofvisual words通过K-means聚类算法对所述特征向量聚类后根据模糊理论进行编码，得到输出向量，并对所述输出向量进行归一化处理，得到文本情感分析模型，包括：

[0029] 从所述训练集中抽取训练样本，通过所述bert base chinese模型进行语义特征提取，对提取出的特征使用K-means聚类的方法，得到聚类中心列表；

[0030] 对提取出的特征使用所述改进Bag ofvisual words进行编码，每个样本被编码为数值向量；

[0031] 将所述特征向量转化为概率值。

[0032] 一种基于迁移学习和改进词袋模型的文本情感分析系统，所述系统包括：

[0033] 数据采集模块，用于采集不同种类商品的各个评论数据，并将各个所述评论数据构建为数据集；

[0034] 预处理模块，用于对所述数据集进行预处理，得到处理后的综合评论数据集；

[0035] 预训练模块，用于根据所述综合评论数据集预训练特征提取器，其中，将bertbase chinese模型作为特征提取器，采用MLM在所述综合评论数据集上进行预训练；

[0036] 特征提取模块，用于构建特定商品评论数据集，将所述特定商品评论数据集输入至所述bert base chinese模型中，提取出特征向量；

[0037] 模型训练模块，用于将所述特征向量输入至改进Bag ofvisual words中，所述改进Bag ofvisual words通过K-means聚类算法对所述特征向量聚类后根据模糊理论进行编码，得到输出向量，并对所述输出向量进行归一化处理，得到文本情感分析模型；

[0038] 情感分析模块,用于通过所述文本情感分析模型进行文本情感分析。

[0039] 一种计算机设备,包括存储器和处理器,所述存储器存储有计算机程序,所述处理器执行所述计算机程序时实现以下步骤:

[0040] 采集不同种类商品的各个评论数据,并将各个所述评论数据构建为数据集;

[0041] 对所述数据集进行预处理,得到处理后的综合评论数据集;

[0042] 根据所述综合评论数据集预训练特征提取器,其中,将bert base chinese模型作为特征提取器,采用MLM在所述综合评论数据集上进行预训练;

[0043] 构建特定商品评论数据集,将所述特定商品评论数据集输入至所述bert base chinese模型中,提取出特征向量;

[0044] 将所述特征向量输入至改进Bag ofvisual words中,所述改进Bag ofvisual words通过K-means聚类算法对所述特征向量聚类后根据模糊理论进行编码,得到输出向量,并对所述输出向量进行归一化处理,得到文本情感分析模型;

[0045] 通过所述文本情感分析模型进行文本情感分析。

[0046] 一种计算机可读存储介质,其上存储有计算机程序,所述计算机程序被处理器执行时实现以下步骤:

[0047] 采集不同种类商品的各个评论数据,并将各个所述评论数据构建为数据集;

[0048] 对所述数据集进行预处理,得到处理后的综合评论数据集;

[0049] 根据所述综合评论数据集预训练特征提取器,其中,将bert base chinese模型作为特征提取器,采用MLM在所述综合评论数据集上进行预训练;

[0050] 构建特定商品评论数据集,将所述特定商品评论数据集输入至所述bert base chinese模型中,提取出特征向量;

[0051] 将所述特征向量输入至改进Bag ofvisual words中,所述改进Bag ofvisual words通过K-means聚类算法对所述特征向量聚类后根据模糊理论进行编码,得到输出向量,并对所述输出向量进行归一化处理,得到文本情感分析模型;

[0052] 通过所述文本情感分析模型进行文本情感分析。

[0053] 上述基于迁移学习和改进词袋模型的文本情感分析方法、系统、计算机设备及存储介质,通过采集不同种类商品的各个评论数据,并将各个所述评论数据构建为数据集;对所述数据集进行预处理,得到处理后的综合评论数据集;根据所述综合评论数据集预训练特征提取器,其中,将bert base chinese模型作为特征提取器,采用MLM在所述综合评论数据集上进行预训练;构建特定商品评论数据集,将所述特定商品评论数据集输入至所述bert base chinese模型中,提取出特征向量;将所述特征向量输入至改进Bag ofvisual words中,所述改进Bag of visual words通过K-means聚类算法对所述特征向量聚类后根据模糊理论进行编码,得到输出向量,并对所述输出向量进行归一化处理,得到文本情感分析模型;通过所述文本情感分析模型进行文本情感分析。通过迁移学习和Bag ofvisual words方法,能够很好的处理不断涌现的新类别商品的评论,同时在重新训练模型时,由于需要学习的参数较少,所以不仅能够减小计算的成本,而且也能成功克服小数据集的限制;此外,不需要再对特征提取器进行微调,只需要根据训练数据更新聚类中心即可,这种训练策略,在学习新知识的同时很好的保留模型以前学过的知识,减少“灾难性遗忘”问题,降低文本情感分析的成本。

附图说明

[0054] 图1为一个实施例中基于迁移学习和改进词袋模型的文本情感分析方法的应用环境图；

[0055] 图2为一个实施例中基于迁移学习和改进词袋模型的文本情感分析方法的流程示意图；

[0056] 图3为一个实施例中训练文本情感分析器的过程示意图；

[0057] 图4为一个实施例中搭建文本情感分析模型的过程示意图；

[0058] 图5为一个实施例中基于迁移学习和改进词袋模型的文本情感分析系统的结构框图；

[0059] 图6为一个实施例中计算机设备的内部结构图。

具体实施方式

[0060] 为了使本申请的目的、技术方案及优点更加清楚明白，以下结合附图及实施例，对本申请进行进一步详细说明。应当理解，此处描述的具体实施例仅仅用以解释本申请，并不用于限定本申请。

[0061] 可以理解，本申请所使用的术语“第一”、“第二”等可在本文中用于描述阈值数量，但这些阈值数量不受这些术语限制。这些术语仅用于将第一个阈值数量与另一个阈值数量区分。举例来说，在不脱离本申请的范围的情况下，可以将第一阈值数量称为第二阈值数量，且类似地，可将第二阈值数量称为第一阈值数量。第一阈值数量和第二阈值数量两者都是阈值数量，但其不是同一阈值数量。

[0062] 本申请实施例提供的基于迁移学习和改进词袋模型的文本情感分析方法，可以应用于如图1所示的应用环境中。如图1所示，该应用环境包括计算机设备110。计算机设备110可以采集不同种类商品的各个评论数据，并将各个评论数据构建为数据集；计算机设备110可以对数据集进行预处理，得到处理后的综合评论数据集；计算机设备110可以根据综合评论数据集预训练特征提取器，其中，将bert base chinese模型作为特征提取器，采用MLM在综合评论数据集上进行预训练；计算机设备110可以构建特定商品评论数据集，将特定商品评论数据集输入至bert base chinese模型中，提取出特征向量；计算机设备110可以将特征向量输入至改进Bag ofvisual words中，改进Bag ofvisual words通过K-means聚类算法对特征向量聚类后根据模糊理论进行编码，得到输出向量，并对输出向量进行归一化处理，得到文本情感分析模型；计算机设备110可以通过文本情感分析模型进行文本情感分析。其中，计算机设备110可以但不限于是各种个人计算机、笔记本电脑、智能手机、机器人、无人飞行器、平板电脑等设备。

[0063] 在一个实施例中，如图2所示，提供了一种基于迁移学习和改进词袋模型的文本情感分析方法，包括以下步骤：

[0064] 步骤202，采集不同种类商品的各个评论数据，并将各个评论数据构建为数据集。

[0065] 计算机设备可以采集不同种类商品的各个评论数据，在本实施例中，采集的各个评论数据可以包含有10种类别的商品评论数据，商品评论数据可以有6万多条，包含有正向评论和负向评论各3万条左右。举例说明，采集的商品评论数据可以分别包括书籍(3851条)、平板(10000条)、手机(2323条)、水果(10000条)、洗发水(10000条)、热水器(575条)、蒙

牛(2033条)、衣服(10000条)、计算机(3992条)、酒店(10000条)。计算机设备采集到各个评论数据后,可以构建数据集。

[0066] 步骤204,对数据集进行预处理,得到处理后的综合评论数据集。

[0067] 其中,预处理可以是针对预训练特征提取器做的操作。预训练特征提取器至需要用到数据集中的评论部分,因此处理后的综合评论数据集中仅包含有评论部分。

[0068] 步骤206,根据综合评论数据集预训练特征提取器,其中,将bert base chinese模型作为特征提取器,采用MLM在综合评论数据集上进行预训练。

[0069] 计算机设备可以使用由Hugging Face提供的bert base chinese模型作为特征提取器,其中,bert base chinese模型已在大型的中文语料库中进行预训练,其中大约有1.1亿个参数需要学习,所以可想而知,如果不使用迁移学习的方法,那每次重新训练一个中文情感分析模型,需要花费的计算成本是巨大的。具体的,bertbase chinese特征提取器已在大型通用语料库中预训练过了,此处为了提高特征提取器的效果,所以进一步在综合评论数据集上预训练,采用的预训练方法是MLM。

[0070] 步骤208,构建特定商品评论数据集,将特定商品评论数据集输入至bert base chinese模型中,提取出特征向量。

[0071] 特征提取器进一步预训练成功后,计算机设备中可以再构造特定商品评论数据集。其中,特定商品评论数据集与综合评论数据集是不一样的,特定商品评论数据集可以用来提取特征向量。在本实施例中,在训练特定商品评论的情感分析模型时,把特定商品评论数据集输入到进一步预训练的bert base chinese模型中提取特征,在训练过程中bert base chinese模型的参数不需要再学习,即迁移学习。

[0072] 步骤210,将特征向量输入至改进Bag ofvisual words中,改进Bag ofvisual words通过K-means聚类算法对特征向量聚类后根据模糊理论进行编码,得到输出向量,并对输出向量进行归一化处理,得到文本情感分析模型。

[0073] 传统的词袋模型Bag ofvisual words包含有特征提取、k-means聚类获得聚类中心、编码、归一化。而改进词袋模型即改进Bag ofvisual words在编码时,根据模糊理论进行编码。

[0074] 步骤212,通过文本情感分析模型进行文本情感分析。

[0075] 在本实施例中,通过迁移学习和Bag ofvisual words方法,能够很好的处理不断涌现的新类别商品的评论,同时在重新训练模型时,由于需要学习的参数较少,所以不仅能够减小计算的成本,而且也能成功克服小数据集的限制;此外,不需要再对特征提取器进行微调,只需要根据训练数据更新聚类中心即可,这种训练策略,在学习新知识的同时很好的保留模型以前学过的知识,减少“灾难性遗忘”问题,降低文本情感分析的成本。

[0076] 在一个实施例中,提供一种基于迁移学习和改进词袋模型的文本情感分析方法还可以包括构建数据集的过程,具体过程包括:将各个评论数据以csv的形式保存,且每条数据包含有类别、正负标签、评论。

[0077] 其中,csv的形式保存的评论数据中每一行是一条商品评论数据,每条数据包含三部分内容,分别是类别、正负标签、评论。

[0078] 在一个实施例中,提供一种基于迁移学习和改进词袋模型的文本情感分析方法还可以包括数据预处理的过程,具体过程包括:从数据集中取出各个评论数据的评论部分;使

用正则表达式的方式去掉各个评论部分中的无意义符号以及非中文内容,得到综合评论数据集。

[0079] 在一个实施例中,提供一种基于迁移学习和改进词袋模型的文本情感分析方法还可以包括预训练特征提取器的过程,具体过程包括:通过Tokenizer工具对输入的特定商品评论数据集数据集中的数据进行分词,并在分词后的样本上加上Token;获取bert base chinese模型在预训练时的字典,并根据字典将各个Token映射为对应的ID;通过bertbase chinese模型将特定商品评论数据集中映射为ID的等长样本转化为数值矩阵,并提取特定商品评论数据集中句子的语义特征和Token的上下文信息,经过输出层输出。

[0080] Hugging Face提供了一个名为Tokenizer的工具,可以对输入的中文评论数据按字为单位进行分词,并在分词后的样本上加上特殊的token,再根据bert base chinese模型在预训练时获得的字典把每个token映射为对应的id。其中,token是指文本进行分割后的最小单元,在本实施例中可以为字。

[0081] 其中,[CLS]放在句子的首位,经过特征提取器得到的表征向量C可以用于后续的分类任务;[SEP]标志放在句子的结尾;[UNK]指的是未知字符;[MASK]用于遮盖句子中的一些单词,将单词用[MASK]遮盖之后,再利用bert base chinese模型输出的[MASK]向量预测单词是什么,这也正是模型预训练的任务之一;[PAD]用来填充小于最大长度的句子,使得数据等长输入模型。

[0082] 在本实施例中,bert base chinese模型由三部分组成:嵌入层(Embedding Layer)、Transformer的编码器和输出层。其中,嵌入层用于把映射为id的等长样本转化为[512,768]维度的数值矩阵表示;Transformer的编码器用于提取输入句子的语义特征和token的上下文信息,是一个动态的编码器;输出层对编码器的输出进行处理,以完成不同的下游任务。

[0083] 在一个实施例中,提供一种基于迁移学习和改进词袋模型的文本情感分析方法还可以包括使用MLM任务进行预训练的过程,具体过程包括:从加上Token的样本中选取目标占比的目标Token;选取第一阈值数量的目标Token替换为mask,选取第二阈值数量的目标Token替换为随机Token,选取第三阈值数量的目标Token保留。

[0084] 在对bert base chinese模型进一步预训练时,由于对中文文本情感分析并没有处理句子对的情况,所以只需要使用MLM任务进一步预训练模型的语义理解能力。其中,目标占比可以是15%;第一阈值数量可以是80%;第二阈值数量可以是10%;第三阈值数量可以是10%。

[0085] 具体的,从加上Token的样本中选取15%的token之后,并不是所有的都替换成[mask]标记符。实际操作是:从这选出的15%部分中,将其中的80%替换成[mask];10%替换成一个随机的token;剩下的10%保留原来的token。其中,用mask_token_list列表来保存被[mask]替换掉的原token,用mask_position_list列表来保存被[mask]替换掉的原token在样本中的位置。

[0086] 在一个实施例中,提供一种基于迁移学习和改进词袋模型的文本情感分析方法还可以包括特定商品评论数据集进行处理的过程,具体过程包括:获取特定商品的评论数据并构建初步特定商品评论数据集;对初步特定商品评论数据集中的评论数据用正则表达式的方式进行预处理,得到处理后的特定商品评论数据集;对特定商品评论数据集进行划分,

构建训练集、验证集、测试集。

[0087] 在本实施例中，特定商品可以是数码类和零食类两类，数码类商品的评论数据集有4000条评论，正、负向评论各约2000条，零食商品的评论数据集有5000条评论，正、负向评论各约2500条。对两个数据集中的评论数据用正则表达式的方法去除无意义符号和非中文内容，再分别按照80％、10％、10％的比例，对两个数据集进行划分，构建训练集、验证集和测试集。

[0088] 在一个实施例中，提供的一种基于迁移学习和改进词袋模型的文本情感分析方法还可以包括改进Bag ofvisual words的处理过程，具体过程包括：从训练集中抽取训练样本，通过bert base chinese模型进行语义特征提取，对提取出的特征使用K-means聚类的方法，得到聚类中心列表；对提取出的特征使用改进Bag ofvisual words进行编码，每个样本被编码为数值向量；将特征向量转化为概率值。

[0089] 其中，在训练一类商品评论的情感分析器时，从该类商品的评论数据集的训练集中，随机选择50％的样本，由进一步预训练的bertbase chinese模型对这些样本进行语义特征提取。

[0090] 提取完特征后，采取K-means聚类算法对这些特征向量进行聚类。聚类完成后，得到K个聚类中心，把这些聚类中心向量保存在Centre_List列表中，这里的K等于300，Centre_List列表的长度也为300。

[0091] 编码时，根据模糊理论，局部特征在与各个聚类中心计算完欧式距离后，不再只取0和1进行编码，而是根据公式$m(D_i,C_j)=\exp(-(D(i,j)-\min(D))^2/\sigma)$，使用(0,1]之间的数编码。这也正是对传统Bag ofvisual words方法改进的核心和该方法能应用于自然语言处理并取得良好效果的关键。公式$m(D_i,C_j)=\exp(-(D(i,j)-\min(D))^2/\sigma)$中，$m(D_i,C_j)$表示样本的第i个局部特征和在第j个位置的编码；D(i,j)表示样本的第i个局部特征与第j个聚类中心欧式距离；σ为超参数，可以根据模型的效果调整；min(D)表示样本的第i个局部特征与所有聚类中心的最短欧式距离。

[0092] 得到样本第i个局部特征的所有位置的编码，就可以把该局部特征表示为300维的数值向量，用$A(D_i,C_k)$表示，即样本的第i个局部特征和在第k个位置的数值向量。

[0093] 得到一个样本的所有局部特征的向量表示后，可以通过公式$D_p=\sum_{i=1}^{n}A(D_i,C_k)$把这些向量加起来，得到样本的输出$D_p$，其中n是局部特征的数量。

[0094] 编码完成后，一个样本得到300维的向量表示，用softmax函数对输出值进行归一化操作，把向量中所有与的值都转化为概率(0～1之间)值，所有概率值加起来等于1。

[0095] 在一个实施例中，训练文本情感分析器的过程如图3所示，通过构建电商平台综合中文评论数据集，接着进行数据预处理，再进一步预训练特征提取器；接着，可以进行特定商品评论数据集构建、预处理、划分，构成训练集、验证集、测试集；其中，训练集可以用于后续文本情感分析过程中的特征提取，以及改进Bag ofvisual words编码；初始化模型中需要训练的参数后，可以根据训练集、验证集中的数据训练模型，并进行模型验证后保存最优参数模型，然后通过测试集中的数据测试模型，最终得到文本情感分析器。

[0096] 在一个实施例中，提供的一种基于迁移学习和改进词袋模型的文本情感分析方法还可以包括搭建文本情感分析模型的过程，如图4所示，训练集中随机抽取50％的样本，用bert base chinese模型进行语义特征提取，对提取出的特征用K-means聚类的方法，得到

聚类中心列表；

[0097] 接着，对训练集中的所有样本用bert base chinese模型进行语义特征提取，对提取出的特征用改进Bag ofvisual words方法进行编码，每个样本被编码为一个300维的数值向量；

[0098] 接着，全连接层把输入的包含所有特征信息的数值向量，转化为最终分类成各个类别的概率，此处是一个二分类任务，使用softmax函数作为激活函数，全连接层的输出包含两个神经元。在训练一个新模型时，只有此处全连接层的参数需要学习。

[0099] 其中，要控制参数不学习，pytorch中关于网络的反向传播操作是基于Variable对象，Variable中有一个参数requires_grad，将requires_grad＝False，网络就不会对该层计算梯度。验证模型和测试模型时，直接使用训练时产生的聚类中心，不用再K-means聚类。

[0100] 应该理解的是，虽然上述流程图中的各个步骤按照箭头的指示依次显示，但是这些步骤并不是必然按照箭头指示的顺序依次执行。除非本文中有明确的说明，这些步骤的执行并没有严格的顺序限制，这些步骤可以以其它的顺序执行。而且，上述流程图中的至少一部分步骤可以包括多个子步骤或者多个阶段，这些子步骤或者阶段并不必然是在同一时刻执行完成，而是可以在不同的时刻执行，这些子步骤或者阶段的执行顺序也不必然是依次进行，而是可以与其它步骤或者其它步骤的子步骤或者阶段的至少一部分轮流或者交替地执行。

[0101] 在一个实施例中，如图5所示，提供了一种基于迁移学习和改进词袋模型的文本情感分析系统，包括：数据采集模块510、预处理模块520、预训练模块530、特征提取模块540、模型训练模块550和情感分析模块560，其中：

[0102] 数据采集模块510，用于采集不同种类商品的各个评论数据，并将各个评论数据构建为数据集；

[0103] 预处理模块520，用于对数据集进行预处理，得到处理后的综合评论数据集；

[0104] 预训练模块530，用于根据综合评论数据集预训练特征提取器，其中，将bert base chinese模型作为特征提取器，采用MLM在综合评论数据集上进行预训练；

[0105] 特征提取模块540，用于构建特定商品评论数据集，将特定商品评论数据集输入至bert base chinese模型中，提取出特征向量；

[0106] 模型训练模块550，用于将特征向量输入至改进Bag ofvisual words中，改进Bag ofvisual words通过K-means聚类算法对特征向量聚类后根据模糊理论进行编码，得到输出向量，并对输出向量进行归一化处理，得到文本情感分析模型；

[0107] 情感分析模块560，用于通过文本情感分析模型进行文本情感分析。

[0108] 在一个实施例中，数据采集模块510还用于将各个评论数据以csv的形式保存，且每条数据包含有类别、正负标签、评论。

[0109] 在一个实施例中，预处理模块520还用于从数据集中取出各个评论数据的评论部分；使用正则表达式的方式去掉各个评论部分中的无意义符号以及非中文内容，得到综合评论数据集。

[0110] 在一个实施例中，特征提取模块540还用于通过Tokenizer工具对输入的目标数据集中的数据进行分词，并在分词后的样本上加上Token；获取bert base chinese模型在预训练时的字典，并根据字典将各个Token映射为对应的ID；通过bertbase chinese模型将特

定商品评论数据集中映射为ID的等长样本转化为数值矩阵,并提取特定商品评论数据集中句子的语义特征和Token的上下文信息,经过输出层输出。

[0111] 在一个实施例中,模型训练模块550还用于使用MLM任务对bert base chinese模型进行预训练;从加上Token的样本中选取目标占比的目标Token;选取第一阈值数量的目标Token替换为mask,选取第二阈值数量的目标Token替换为随机Token,选取第三阈值数量的目标Token保留。

[0112] 在一个实施例中,数据采集模块510还用于获取特定商品的评论数据并构建初步特定商品评论数据集;对初步特定商品评论数据集中的评论数据用正则表达式的方式进行预处理,得到处理后的特定商品评论数据集;对特定商品评论数据集进行划分,构建训练集、验证集、测试集。

[0113] 在一个实施例中,模型训练模块550还用于从训练集中抽取训练样本,通过bert base chinese模型进行语义特征提取,对提取出的特征使用K-means聚类的方法,得到聚类中心列表;对提取出的特征使用改进Bag ofvisual words进行编码,每个样本被编码为数值向量;将特征向量转化为概率值。

[0114] 在一个实施例中,提供了一种计算机设备,该计算机设备可以是终端,其内部结构图可以如图6所示。该计算机设备包括通过系统总线连接的处理器、存储器、网络接口、显示屏和输入装置。其中,该计算机设备的处理器用于提供计算和控制能力。该计算机设备的存储器包括非易失性存储介质、内存储器。该非易失性存储介质存储有操作系统和计算机程序。该内存储器为非易失性存储介质中的操作系统和计算机程序的运行提供环境。该计算机设备的网络接口用于与外部的终端通过网络连接通信。该计算机程序被处理器执行时以实现一种基于迁移学习和改进词袋模型的文本情感分析方法。该计算机设备的显示屏可以是液晶显示屏或者电子墨水显示屏,该计算机设备的输入装置可以是显示屏上覆盖的触摸层,也可以是计算机设备外壳上设置的按键、轨迹球或触控板,还可以是外接的键盘、触控板或鼠标等。

[0115] 本领域技术人员可以理解,图6中示出的结构,仅仅是与本申请方案相关的部分结构的框图,并不构成对本申请方案所应用于其上的计算机设备的限定,具体的计算机设备可以包括比图中所示更多或更少的部件,或者组合某些部件,或者具有不同的部件布置。

[0116] 在一个实施例中,提供了一种计算机设备,包括存储器和处理器,存储器中存储有计算机程序,该处理器执行计算机程序时实现以下步骤:

[0117] 采集不同种类商品的各个评论数据,并将各个评论数据构建为数据集;

[0118] 对数据集进行预处理,得到处理后的综合评论数据集;

[0119] 根据综合评论数据集预训练特征提取器,其中,将bert base chinese模型作为特征提取器,采用MLM在综合评论数据集上进行预训练;

[0120] 构建特定商品评论数据集,将特定商品评论数据集输入至bert base chinese模型中,提取出特征向量;

[0121] 将特征向量输入至改进Bag ofvisual words中,改进Bag ofvisual words通过K-means聚类算法对特征向量聚类后根据模糊理论进行编码,得到输出向量,并对输出向量进行归一化处理,得到文本情感分析模型;

[0122] 通过文本情感分析模型进行文本情感分析。

[0123] 在一个实施例中，处理器执行计算机程序时还实现以下步骤：将各个评论数据以csv的形式保存，且每条数据包含有类别、正负标签、评论。

[0124] 在一个实施例中，处理器执行计算机程序时还实现以下步骤：从数据集中取出各个评论数据的评论部分；使用正则表达式的方式去掉各个评论部分中的无意义符号以及非中文内容，得到综合评论数据集。

[0125] 在一个实施例中，处理器执行计算机程序时还实现以下步骤：通过Tokenizer工具对输入的特定商品评论数据集中的数据进行分词，并在分词后的样本上加上Token；获取bert base chinese模型在预训练时的字典，并根据字典将各个Token映射为对应的ID；通过bertbase chinese模型将特定商品评论数据集中映射为ID的等长样本转化为数值矩阵，并提取特定商品评论数据集中句子的语义特征和Token的上下文信息，经过输出层输出。

[0126] 在一个实施例中，处理器执行计算机程序时还实现以下步骤：从加上Token的样本中选取目标占比的目标Token；选取第一阈值数量的目标Token替换为mask，选取第二阈值数量的目标Token替换为随机Token，选取第三阈值数量的目标Token保留。

[0127] 在一个实施例中，处理器执行计算机程序时还实现以下步骤：获取特定商品的评论数据并构建初步特定商品评论数据集；对初步特定商品评论数据集中的评论数据用正则表达式的方式进行预处理，得到处理后的特定商品评论数据集；对特定商品评论数据集进行划分，构建训练集、验证集、测试集。

[0128] 在一个实施例中，处理器执行计算机程序时还实现以下步骤：从训练集中抽取训练样本，通过bert base chinese模型进行语义特征提取，对提取出的特征使用K-means聚类的方法，得到聚类中心列表；对提取出的特征使用改进Bag of visual words进行编码，每个样本被编码为数值向量；将特征向量转化为概率值。

[0129] 在一个实施例中，提供了一种计算机可读存储介质，其上存储有计算机程序，计算机程序被处理器执行时实现以下步骤：

[0130] 采集不同种类商品的各个评论数据，并将各个评论数据构建为数据集；

[0131] 对数据集进行预处理，得到处理后的综合评论数据集；

[0132] 根据综合评论数据集预训练特征提取器，其中，将bert base chinese模型作为特征提取器，采用MLM在综合评论数据集上进行预训练；

[0133] 构建特定商品评论数据集，将特定商品评论数据集输入至bert base chinese模型中，提取出特征向量；

[0134] 将特征向量输入至改进Bag ofvisual words中，改进Bag ofvisual words通过K-means聚类算法对特征向量聚类后根据模糊理论进行编码，得到输出向量，并对输出向量进行归一化处理，得到文本情感分析模型；

[0135] 通过文本情感分析模型进行文本情感分析。

[0136] 在一个实施例中，计算机程序被处理器执行时还实现以下步骤：将各个评论数据以csv的形式保存，且每条数据包含有类别、正负标签、评论。

[0137] 在一个实施例中，计算机程序被处理器执行时还实现以下步骤：从数据集中取出各个评论数据的评论部分；使用正则表达式的方式去掉各个评论部分中的无意义符号以及非中文内容，得到综合评论数据集。

[0138] 在一个实施例中，计算机程序被处理器执行时还实现以下步骤：通过Tokenizer工

具对输入的特定商品评论数据集中的数据进行分词,并在分词后的样本上加上Token;获取bert base chinese模型在预训练时的字典,并根据字典将各个Token映射为对应的ID;通过bert base chinese模型将特定商品评论数据集中映射为ID的等长样本转化为数值矩阵,并提取特定商品评论数据集中句子的语义特征和Token的上下文信息,经过输出层输出。

[0139] 在一个实施例中,计算机程序被处理器执行时还实现以下步骤:从加上Token的样本中选取目标占比的目标Token;选取第一阈值数量的目标Token替换为mask,选取第二阈值数量的目标Token替换为随机Token,选取第三阈值数量的目标Token保留。

[0140] 在一个实施例中,计算机程序被处理器执行时还实现以下步骤:获取特定商品的评论数据并构建初步特定商品评论数据集;对初步特定商品评论数据集中的评论数据用正则表达式的方式进行预处理,得到处理后的特定商品评论数据集;对特定商品评论数据集进行划分,构建训练集、验证集、测试集。

[0141] 在一个实施例中,计算机程序被处理器执行时还实现以下步骤:从训练集中抽取训练样本,通过bert base chinese模型进行语义特征提取,对提取出的特征使用K-means聚类的方法,得到聚类中心列表;对提取出的特征使用改进Bag of visual words进行编码,每个样本被编码为数值向量;将特征向量转化为概率值。

[0142] 本领域普通技术人员可以理解实现上述实施例方法中的全部或部分流程,是可以通过计算机程序来指令相关的硬件来完成,所述的计算机程序可存储于一非易失性计算机可读取存储介质中,该计算机程序在执行时,可包括如上述各方法的实施例的流程。其中,本申请所提供的各实施例中所使用的对存储器、存储、数据库或其它介质的任何引用,均可包括非易失性和/或易失性存储器。非易失性存储器可包括只读存储器(ROM)、可编程ROM(PROM)、电可编程ROM(EPROM)、电可擦除可编程ROM(EEPROM)或闪存。易失性存储器可包括随机存取存储器(RAM)或者外部高速缓冲存储器。作为说明而非局限,RAM以多种形式可得,诸如静态RAM(SRAM)、动态RAM(DRAM)、同步DRAM(SDRAM)、双数据率SDRAM(DDRSDRAM)、增强型SDRAM(ESDRAM)、同步链路(Synchlink)DRAM(SLDRAM)、存储器总线(Rambus)直接RAM(RDRAM)、直接存储器总线动态RAM(DRDRAM)、以及存储器总线动态RAM(RDRAM)等。

[0143] 以上实施例的各技术特征可以进行任意的组合,为使描述简洁,未对上述实施例中的各个技术特征所有可能的组合都进行描述,然而,只要这些技术特征的组合不存在矛盾,都应当认为是本说明书记载的范围。

[0144] 以上所述实施例仅表达了本申请的几种实施方式,其描述较为具体和详细,但并不能因此而理解为对发明专利范围的限制。应当指出的是,对于本领域的普通技术人员来说,在不脱离本申请构思的前提下,还可以做出若干变形和改进,这些都属于本申请的保护范围。因此,本申请专利的保护范围应以所附权利要求为准。

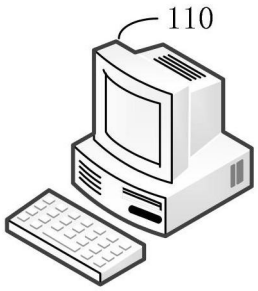

110

图1

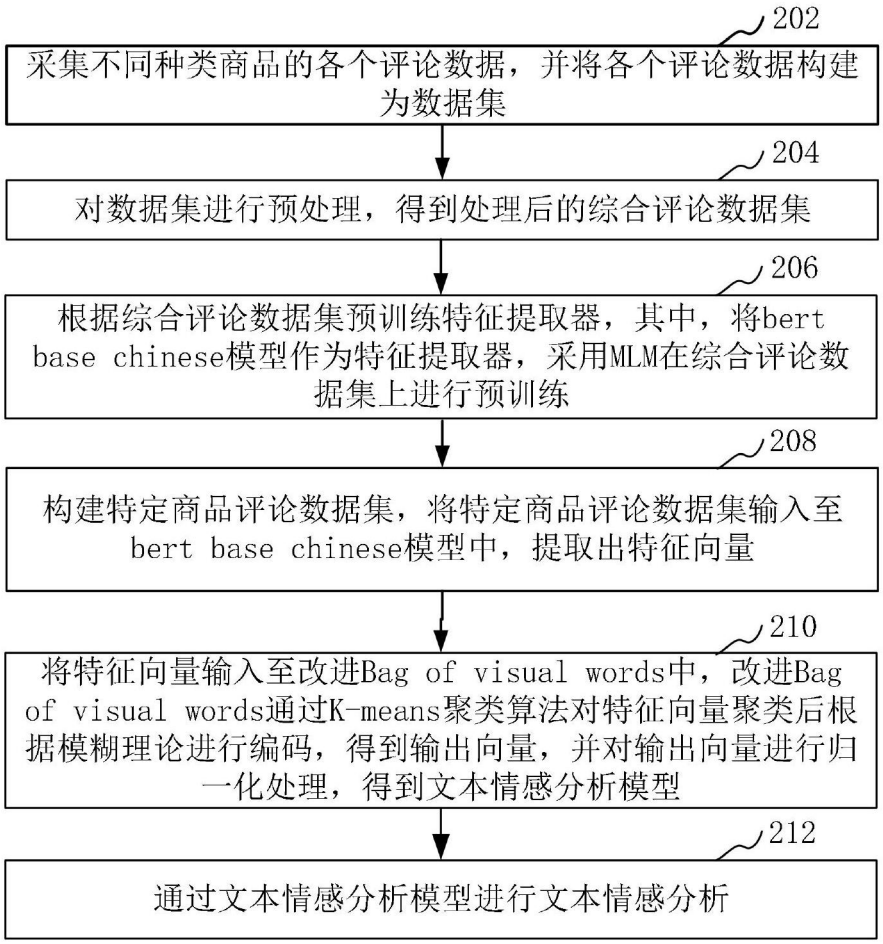

202
采集不同种类商品的各个评论数据，并将各个评论数据构建为数据集
204
对数据集进行预处理，得到处理后的综合评论数据集
206
根据综合评论数据集预训练特征提取器，其中，将bert base chinese模型作为特征提取器，采用MLM在综合评论数据集上进行预训练
208
构建特定商品评论数据集，将特定商品评论数据集输入至bert base chinese模型中，提取出特征向量
210
将特征向量输入至改进Bag of visual words中，改进Bag of visual words通过K-means聚类算法对特征向量聚类后根据模糊理论进行编码，得到输出向量，并对输出向量进行归一化处理，得到文本情感分析模型
212
通过文本情感分析模型进行文本情感分析

图2

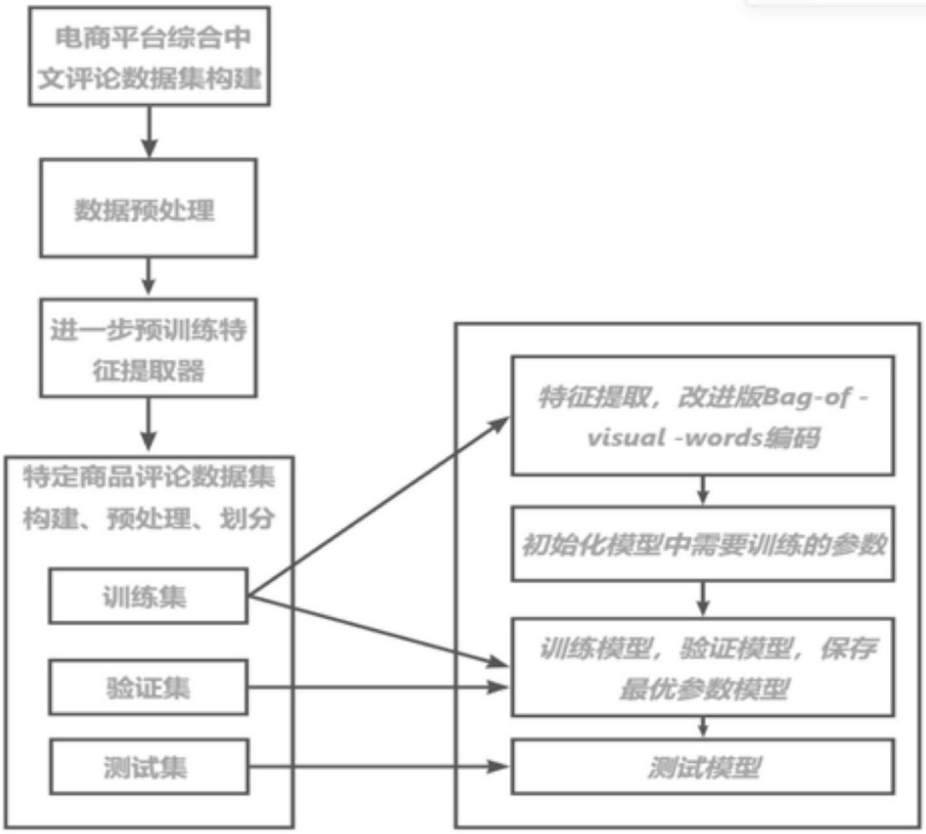

电商平台综合中文评论数据集构建
数据预处理
进一步预训练特征提取器
特定商品评论数据集构建、预处理、划分
训练集
验证集
测试集
特征提取，改进版Bag-of-visual-words编码
初始化模型中需要训练的参数
训练模型，验证模型，保存最优参数模型
测试模型

图3

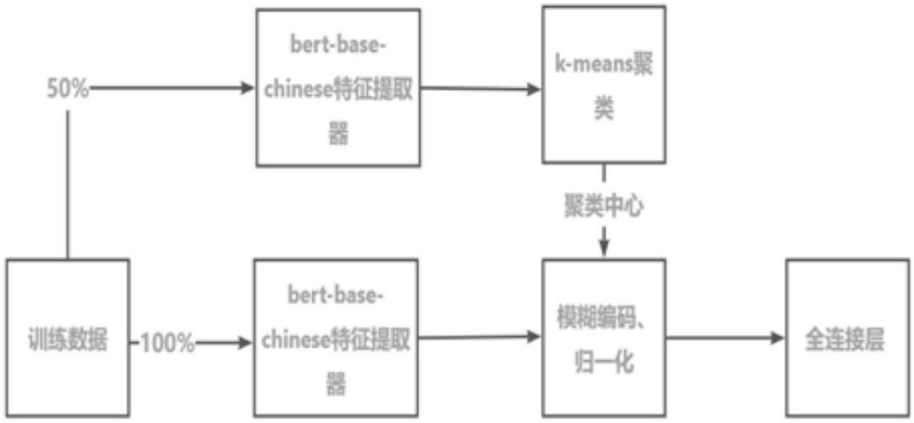

50%
bert-base-chinese特征提取器
k-means聚类
聚类中心
训练数据
100%
bert-base-chinese特征提取器
模糊编码、归一化
全连接层

图4

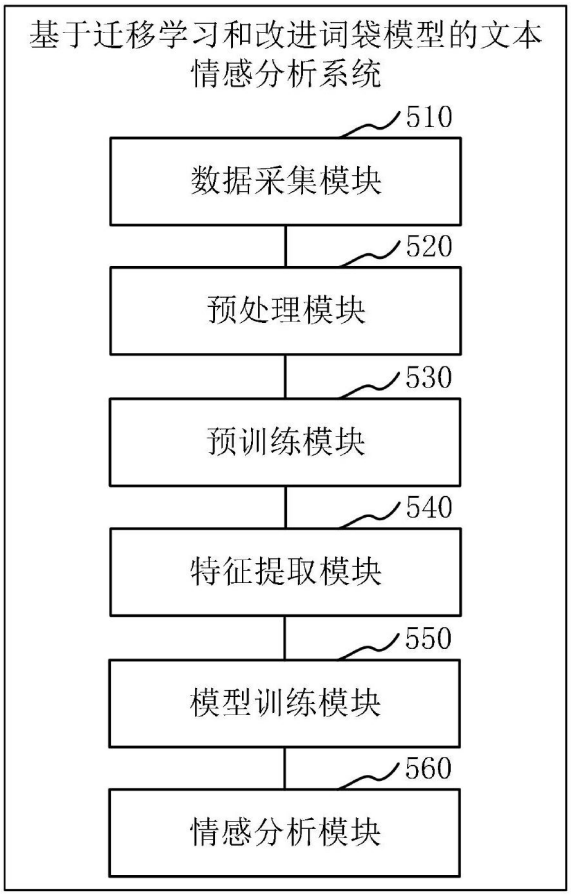

基于迁移学习和改进词袋模型的文本情感分析系统
510
数据采集模块
520
预处理模块
530
预训练模块
540
特征提取模块
550
模型训练模块
560
情感分析模块

图5

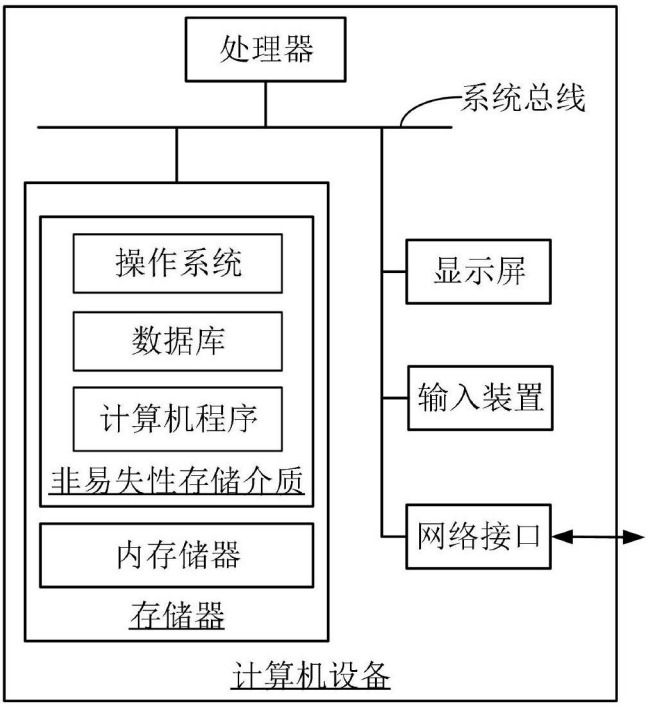

处理器
系统总线
操作系统
数据库
计算机程序
非易失性存储介质
内存储器
存储器
显示屏
输入装置
网络接口
计算机设备

图6