(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 112463964 A
(43)申请公布日 2021.03.09

(21)申请号 202011386332.X

(22)申请日 2020.12.01

(71)申请人 科大讯飞股份有限公司
地址 230088 安徽省合肥市高新区望江西路666号

(72)发明人 葛学志 刘权 陈志刚 王志国 胡国平

(74)专利代理机构 北京集佳知识产权代理有限公司 11227
代理人 付丽

(51)Int.Cl.
G06F 16/35(2019.01)

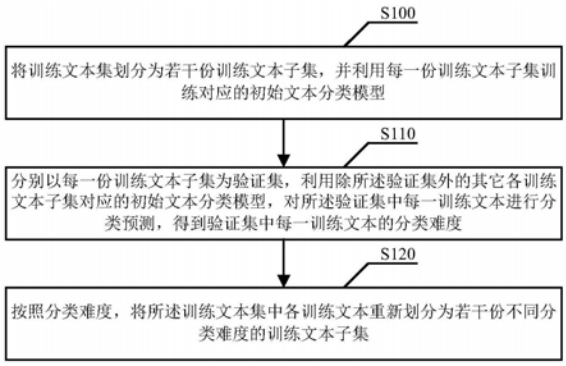
权利要求书3页 说明书13页 附图3页

(54)发明名称

文本分类及模型训练方法、装置、设备及存储介质

(57)摘要

本申请公开了一种文本分类及模型训练方法、装置、设备及存储介质，本申请首先将训练文本集划分为若干份训练文本子集，并利用每一子集训练对应的初始文本分类模型，进而分别以每一子集为验证集，利用除验证集外其它各子集对应的初始文本分类模型对验证集中每一训练文本进行分类预测，以得到每一训练文本的分类难度，按照分类难度，将训练文本集中各训练文本重新划分为若干份不同分类难度的训练文本子集。本申请文本分类方法为更好的训练目标文本分类模型提供了有力的训练数据支撑，后续可以基于不同分类难度的训练文本子集，按照分类难度由低至高顺序递进式训练目标文本分类模型，解决由于训练文本难度不均衡现象导致的模型训练效果不佳的问题。



CN 112463964 A

1.一种文本分类方法,其特征在于,包括:

将训练文本集划分为若干份训练文本子集,并利用每一份训练文本子集训练对应的初始文本分类模型;

分别以每一份训练文本子集为验证集,利用除所述验证集外的其它各训练文本子集对应的初始文本分类模型,对所述验证集中每一训练文本进行分类预测,得到验证集中每一训练文本的分类难度;

按照分类难度,将所述训练文本集中各训练文本重新划分为若干份不同分类难度的训练文本子集。

2.根据权利要求1所述的方法,其特征在于,利用各初始文本分类模型,对验证集中每一训练文本进行分类预测,得到每一训练文本的分类难度的过程,包括:

针对验证集中每一训练文本,利用各初始文本分类模型,分别对所述训练文本进行分类预测,得到每一初始文本分类模型预测的所述训练文本属于标注类别的置信度;

基于各初始文本分类模型预测的所述训练文本属于标注类别的置信度,确定所述训练文本的分类难度。

3.根据权利要求2所述的方法,其特征在于,所述基于各初始文本分类模型预测的所述训练文本属于标注类别的置信度,确定所述训练文本的分类难度,包括:

将各初始文本分类模型预测的所述训练文本属于标注类别的置信度的集合,确定为所述训练文本的分类难度表示;

或,

对各初始文本分类模型预测的所述训练文本属于标注类别的置信度进行数学运算,得到综合置信度,并将所述综合置信度确定为所述训练文本的分类难度表示。

4.根据权利要求1所述的方法,其特征在于,所述按照分类难度,将所述训练文本集中各训练文本重新划分为若干份不同分类难度的训练文本子集,包括:

参考设定的各分类难度区间,确定所述训练文本集中各训练文本的分类难度所处的分类难度区间;

将所述训练文本集中每一训练文本划分至所处分类难度区间对应的新训练文本子集,以得到每一分类难度区间对应的新训练文本子集。

5.根据权利要求1所述的方法,其特征在于,所述按照分类难度,将所述训练文本集中各训练文本重新划分为若干份不同分类难度的训练文本子集,包括:

以分类难度作为聚类条件,对所述训练文本集中各训练文本进行聚类,得到聚类后的若干聚类簇;

将每一聚类簇中的训练文本作为一个新训练文本子集,并基于聚类簇内各训练文本的分类难度确定聚类簇对应的新训练文本子集的分类难度。

6.一种文本分类模型训练方法,其特征在于,包括:

获取权利要求1-5任一项所述的文本分类方法所划分的各不同分类难度的训练文本子集;

按照分类难度由低至高的顺序,依次利用各训练文本子集训练目标文本分类模型。

7.根据权利要求6所述的方法,其特征在于,在所述依次利用各训练文本子集训练目标文本分类模型之后,该方法还包括:

以各不同难度的训练文本子集组成全量训练文本集，并利用所述全量训练文本集训练所述目标文本分类模型。

8.根据权利要求6所述的方法，其特征在于，所述按照分类难度由低至高的顺序，依次利用各训练文本子集训练目标文本分类模型，包括：

按照各训练文本子集的分类难度由低至高顺序，确定与每一训练文本子集对应的训练阶段；

在每一训练阶段中，从所述训练阶段及之前各训练阶段分别对应的训练文本子集中采样训练文本，采样结果组成所述训练阶段对应的训练文本集合；

利用所述训练阶段对应的训练文本集合，训练目标文本分类模型。

9.根据权利要求8所述的方法，其特征在于，所述利用所述训练阶段对应的训练文本集合，训练目标文本分类模型，包括：

利用所述训练阶段对应的训练文本集合，以及所述训练文本集合中各类别文本的分布信息，训练目标文本分类模型。

10.根据权利要求9所述的方法，其特征在于，所述利用所述训练阶段对应的训练文本集合，以及所述训练文本集合中各类别文本的分布信息，训练目标文本分类模型，包括：

将所述训练阶段对应的训练文本集合中每一训练文本的编码特征输入目标文本分类模型；

由所述目标文本分类模型的隐层对所述编码特征进行处理，并基于处理后的编码特征预测所述训练文本在各分类标签上的得分；

由所述目标文本分类模型将所述训练文本在每一分类标签上的得分，与所述训练文本集合中对应分类标签下训练文本的数量相加，得到所述训练文本在每一分类标签上的综合得分，并对各分类标签上的综合得分进行归一化处理；

结合所述训练文本在各分类标签上归一化后的综合得分，以所述训练文本的分类损失为损失函数，训练目标文本分类模型。

11.一种文本分类装置，其特征在于，包括：

初始文本分类模型训练单元，用于将训练文本集划分为若干份训练文本子集，并利用每一份训练文本子集训练对应的初始文本分类模型；

分类难度预测单元，用于分别以每一份训练文本子集为验证集，利用除所述验证集外的其它各训练文本子集对应的初始文本分类模型，对所述验证集中每一训练文本进行分类预测，得到验证集中每一训练文本的分类难度；

训练文本再划分单元，用于按照分类难度，将所述训练文本集中各训练文本重新划分为若干份不同分类难度的训练文本子集。

12.一种文本分类模型训练装置，其特征在于，包括：

训练文本子集获取单元，用于获取权利要求1-5任一项所述的文本分类方法所划分的各不同难度的训练文本子集；

目标文本分类模型训练单元，用于按照分类难度由低至高的顺序，依次利用各训练文本子集训练目标文本分类模型。

13.一种文本分类设备，其特征在于，包括：存储器和处理器；

所述存储器，用于存储程序；

所述处理器，用于执行所述程序，实现如权利要求1～5中任一项所述的文本分类方法的各个步骤。

14.一种存储介质，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，实现如权利要求1～5中任一项所述的文本分类方法的各个步骤。

文本分类及模型训练方法、装置、设备及存储介质

技术领域

[0001] 本申请涉及训练数据处理技术领域，更具体的说，是涉及一种文本分类及模型训练方法、装置、设备及存储介质。

背景技术

[0002] 在自然语言理解领域存在一项基础而重要的任务，即对文本进行分类。为了实现文本分类，现有技术一般通过训练神经网络模型，以通过模型进行文本分类处理。

[0003] 在真实世界的场景下，大量的训练文本常常表现出一种长尾现象(又叫做样本不均衡问题)，不同的训练文本包含的信息量大小不同，这就导致不同的训练文本被模型学习的难易程度不同，也即单个训练文本训练难度不均衡造成的样本不均衡现象。现有技术并未对训练文本进行区分，而是直接基于训练难度不均衡的样本直接训练模型，这就导致模型学习到的知识有限，进而导致最终分类能力不佳。

发明内容

[0004] 鉴于上述问题，提出了本申请以便提供一种文本分类及模型训练方法、装置、设备及存储介质，以解决现有训练文本存在训练难度不均衡现象，在训练模型时会影响模型训练效果的问题。具体方案如下：

[0005] 一种文本分类方法，包括：

[0006] 将训练文本集划分为若干份训练文本子集，并利用每一份训练文本子集训练对应的初始文本分类模型；

[0007] 分别以每一份训练文本子集为验证集，利用除所述验证集外的其它各训练文本子集对应的初始文本分类模型，对所述验证集中每一训练文本进行分类预测，得到验证集中每一训练文本的分类难度；

[0008] 按照分类难度，将所述训练文本集中各训练文本重新划分为若干份不同分类难度的训练文本子集。

[0009] 优选地，所述将训练文本集划分为若干份训练文本子集，包括：

[0010] 采用随机划分的方式，将训练文本集随机划分为若干份互不交叉的训练文本子集。

[0011] 优选地，利用各初始文本分类模型，对验证集中每一训练文本进行分类预测，得到每一训练文本的分类难度的过程，包括：

[0012] 针对验证集中每一训练文本，利用各初始文本分类模型，分别对所述训练文本进行分类预测，得到每一初始文本分类模型预测的所述训练文本属于标注类别的置信度；

[0013] 基于各初始文本分类模型预测的所述训练文本属于标注类别的置信度，确定所述训练文本的分类难度。

[0014] 优选地，所述基于各初始文本分类模型预测的所述训练文本属于标注类别的置信度，确定所述训练文本的分类难度，包括：

[0015] 将各初始文本分类模型预测的所述训练文本属于标注类别的置信度的集合，确定为所述训练文本的分类难度表示；

[0016] 或，

[0017] 对各初始文本分类模型预测的所述训练文本属于标注类别的置信度进行数学运算，得到综合置信度，并将所述综合置信度确定为所述训练文本的分类难度表示。

[0018] 优选地，所述按照分类难度，将所述训练文本集中各训练文本重新划分为若干份不同分类难度的训练文本子集，包括：

[0019] 参考设定的各分类难度区间，确定所述训练文本集中各训练文本的分类难度所处的分类难度区间；

[0020] 将所述训练文本集中每一训练文本划分至所处分类难度区间对应的新训练文本子集，以得到每一分类难度区间对应的新训练文本子集。

[0021] 优选地，所述按照分类难度，将所述训练文本集中各训练文本重新划分为若干份不同分类难度的训练文本子集，包括：

[0022] 以分类难度作为聚类条件，对所述训练文本集中各训练文本进行聚类，得到聚类后的若干聚类簇；

[0023] 将每一聚类簇中的训练文本作为一个新训练文本子集，并基于聚类簇内各训练文本的分类难度确定聚类簇对应的新训练文本子集的分类难度。

[0024] 一种文本分类模型训练方法，包括：

[0025] 获取上述的文本分类方法所划分的各不同分类难度的训练文本子集；

[0026] 按照分类难度由低至高的顺序，依次利用各训练文本子集训练目标文本分类模型。

[0027] 优选地，所述按照分类难度由低至高的顺序，依次利用各训练文本子集训练目标文本分类模型，包括：

[0028] 按照分类难度由低至高的顺序，采用增量训练方式依次利用各训练文本子集训练目标文本分类模型。

[0029] 优选地，在所述依次利用各训练文本子集训练目标文本分类模型之后，该方法还包括：

[0030] 以各不同难度的训练文本子集组成全量训练文本集，并利用所述全量训练文本集训练所述目标文本分类模型。

[0031] 优选地，所述按照分类难度由低至高的顺序，采用增量训练方式依次利用各训练文本子集训练目标文本分类模型，包括：

[0032] 按照各训练文本子集的分类难度由低至高顺序，确定与每一训练文本子集对应的训练阶段；

[0033] 在每一训练阶段中，从所述训练阶段及之前各训练阶段分别对应的训练文本子集中采样训练文本，采样结果组成所述训练阶段对应的训练文本集合；

[0034] 利用所述训练阶段对应的训练文本集合，训练目标文本分类模型。

[0035] 优选地，所述利用所述训练阶段对应的训练文本集合，训练目标文本分类模型，包括：

[0036] 利用所述训练阶段对应的训练文本集合，以及所述训练文本集合中各类别文本的

分布信息，训练目标文本分类模型。

[0037] 优选地，所述利用所述训练阶段对应的训练文本集合，以及所述训练文本集合中各类别文本的分布信息，训练目标文本分类模型，包括：

[0038] 将所述训练阶段对应的训练文本集合中每一训练文本的编码特征输入目标文本分类模型；

[0039] 由所述目标文本分类模型的隐层对所述编码特征进行处理，并基于处理后的编码特征预测所述训练文本在各分类标签上的得分；

[0040] 由所述目标文本分类模型将所述训练文本在每一分类标签上的得分，与所述训练文本集合中对应分类标签下训练文本的数量相加，得到所述训练文本在每一分类标签上的综合得分，并对各分类标签上的综合得分进行归一化处理；

[0041] 结合所述训练文本在各分类标签上归一化后的综合得分，以所述训练文本的分类损失为损失函数，训练目标文本分类模型。

[0042] 一种文本分类装置，包括：

[0043] 初始文本分类模型训练单元，用于将训练文本集划分为若干份训练文本子集，并利用每一份训练文本子集训练对应的初始文本分类模型；

[0044] 分类难度预测单元，用于分别以每一份训练文本子集为验证集，利用除所述验证集外的其它各训练文本子集对应的初始文本分类模型，对所述验证集中每一训练文本进行分类预测，得到验证集中每一训练文本的分类难度；

[0045] 训练文本再划分单元，用于按照分类难度，将所述训练文本集中各训练文本重新划分为若干份不同分类难度的训练文本子集。

[0046] 一种文本分类模型训练装置，包括：

[0047] 训练文本子集获取单元，用于获取上述的文本分类方法所划分的各不同难度的训练文本子集；

[0048] 目标文本分类模型训练单元，用于按照分类难度由低至高的顺序，依次利用各训练文本子集训练目标文本分类模型。

[0049] 一种文本分类设备，包括：存储器和处理器；

[0050] 所述存储器，用于存储程序；

[0051] 所述处理器，用于执行所述程序，实现如上所述的文本分类方法的各个步骤。

[0052] 一种存储介质，其上存储有计算机程序，所述计算机程序被处理器执行时，实现如上所述的文本分类方法的各个步骤。

[0053] 借由上述技术方案，本申请的文本分类方法首先将训练文本集划分为若干份训练文本子集，并利用每一子集训练对应的初始文本分类模型，进而分别以每一子集为验证集，利用除验证集外其它各子集对应的初始文本分类模型对验证集中每一训练文本进行分类预测，以得到每一训练文本的分类难度，该分类难度衡量了训练文本被模型学习的难易程度，进而可以按照分类难度，将训练文本集中各训练文本重新划分为若干份不同分类难度的训练文本子集，后续可以基于不同分类难度的训练文本子集，按照分类难度由低至高顺序递进式训练目标文本分类模型，从而解决由于训练文本难度不均衡现象导致的模型训练效果不佳的问题。本申请实施例提供的文本分类方法能够为目标文本分类模型的训练提供按照分类难度划分后的训练文本子集，也即为更好的训练目标文本分类模型提供了有力的

训练数据支撑。同时，按照文本分类方法所提供的训练文本，可以解决由于训练文本难度不均衡现象导致的模型训练效果不佳的问题，使得训练后的目标文本分类模型分类能力更加优秀。

附图说明

[0054] 通过阅读下文优选实施方式的详细描述，各种其他的优点和益处对于本领域普通技术人员将变得清楚明了。附图仅用于示出优选实施方式的目的，而并不认为是对本申请的限制。而且在整个附图中，用相同的参考符号表示相同的部件。在附图中：

[0055] 图1为本申请实施例提供的文本分类方法的一流程示意图；

[0056] 图2示例了一种目标文本分类模型的训练过程示意图；

[0057] 图3为本申请实施例提供的一种文本分类装置结构示意图；

[0058] 图4为本申请实施例提供的一种文本分类模型训练装置结构示意图；

[0059] 图5为本申请实施例提供的文本分类设备的结构示意图。

具体实施方式

[0060] 下面将结合本申请实施例中的附图，对本申请实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本申请一部分实施例，而不是全部的实施例。基于本申请中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本申请保护的范围。

[0061] 本申请提供了一种文本分类方案，能够对分类难度不均衡的训练文本集进行划分，得到不同分类难度的训练文本子集。在此基础上，进一步提供了一种模型训练方案，应用该不同分类难度的训练文本子集，按照分类难度由低至高顺序，依次训练目标文本分类模型，使得训练后的目标文本分类模型的分类效果更佳。

[0062] 本申请方案可以基于具备数据处理能力的终端实现，该终端可以是手机、电脑、服务器、云端等。

[0063] 接下来，结合图1所述，本申请的文本分类方法可以包括如下步骤：

[0064] 步骤S100、将训练文本集划分为若干份训练文本子集，并利用每一份训练文本子集训练对应的初始文本分类模型。

[0065] 具体的，可以预先规定划分后训练文本子集的数量k，则可以采用随机划分或其它方式，将训练文本集随机划分为k份训练文本子集。k可以是大于等于2的设定数值。

[0066] 除此之外，还可以采用其它划分方式，将训练文本集划分为若干份训练文本子集。各个训练文本子集互不交叉，也即各个训练文本子集中包含的训练文本互不相同。各个训练文本子集与训练文本集服从相同的样本空间分布。

[0067] 针对划分后的每一份训练文本子集，利用训练文本子集训练对应的初始文本分类模型。

[0068] 可以理解的是，若训练文本子集有k份，分别表示为x_i(i＝1～k)，则对应的初始文本分类模型也有k个，与k份训练文本子集一一对应，分别表示为y_i(i＝1～k)。

[0069] 对于初始文本分类模型，其可以采用多种不同架构的神经网络模型，文本分类模型用于实现文本分类，具体的分类任务可以是多种，如阅读理解、序列标注等。

[0070] 步骤S110、分别以每一份训练文本子集为验证集，利用除所述验证集外的其它各训练文本子集对应的初始文本分类模型，对所述验证集中每一训练文本进行分类预测，得到验证集中每一训练文本的分类难度。

[0071] 具体的，假设以x_j为验证集，则可以利用$y_i(i\neq j)$分别对x_j中每一训练文本进行分类预测，以得到验证集x_j中每一训练文本的分类难度。

[0072] 通过j遍历1-k，也即分别以各份训练文本子集作为验证集，最终可以得到训练文本集中每一训练文本的分类难度。

[0073] 本步骤中，针对每一训练文本，使用k-1个初始文本分类模型进行分类预测，可以减小单个模型预测造成的偏差，从而提高模型结果的置信度，降低个体决策的失误率。另外，由于每一初始文本分类模型虽然是通过不同训练文本子集训练出来的，但是这些训练文本子集都是来自原始的训练文本集，它们都是同一个类别标签下的不同说法或者表达而已，具有很高的类间相似度，因此，通过综合k-1个初始文本分类模型对训练文本进行分类预测，结合分类预测结果可以衡量训练文本的分类难度。

[0074] 分类难度衡量了训练文本被模型学习的难易程度，也即反映了训练文本包含的信息量多少。示例如，若多个初始分类模型对一训练文本的预测分类结果的置信度均较高，则说明该训练文本包含的信息量较少，学习难度较低，当前初始分类模型可以轻松学习到该训练文本所包含的信息。反之，若多个初始分类模型对一训练文本的预测分类结果的置信度均较低，则说明该训练文本包含的信息量较多，学习难度较高，当前初始分类模型无法学习到该训练文本所包含的全部信息。

[0075] 步骤S120、按照分类难度，将所述训练文本集中各训练文本重新划分为若干份不同分类难度的训练文本子集。

[0076] 具体的，在确定了每一训练文本的分类难度之后，可以按照各训练文本的分类难度，将训练文本集中各训练文本重新划分为若干份不同分类难度的训练文本子集，划分后的各训练文本子集对应的分类难度不同。

[0077] 本申请实施例提供的文本分类方法，能够确定出训练文本集中各训练文本的分类难度，进而可以按照分类难度，将训练文本集中各训练文本重新划分为若干份不同分类难度的训练文本子集，后续可以基于不同分类难度的训练文本子集，按照分类难度由低至高顺序递进式训练目标文本分类模型，这种渐进式的学习本质上模拟人脑学习新知识的过程，即由易到难、循序渐进，从而解决由于训练文本难度不均衡现象导致的模型训练效果不佳的问题。本申请实施例提供的文本分类方法能够为目标文本分类模型的训练提供按照分类难度划分后的训练文本子集，也即为更好的训练目标文本分类模型提供了有力的训练数据支撑。

[0078] 在本申请的一些实施例中，对上述步骤S110中，利用各初始文本分类模型，对验证集中每一训练文本进行分类预测，得到每一训练文本的分类难度的过程进行介绍。

[0079] 具体的，针对验证集中每一训练文本，利用各初始文本分类模型，分别对所述训练文本进行分类预测，得到每一初始文本分类模型预测的所述训练文本属于标注类别的置信度。

[0080] 以上述实施例中示例的总共有k个初始文本分类模型为例，则针对一验证集中一训练文本，可以利用其中k-1个初始文本分类模型分别对其进行分类预测，总共可以得到k-

1个置信度。

[0081] 其中，该置信度为模型预测的训练文本属于标注类别的置信度，也即预测的训练文本在所属标签对应的类别上的置信度得分。

[0082] 进一步的，基于各初始文本分类模型预测的所述训练文本属于标注类别的置信度，确定所述训练文本的分类难度。

[0083] 具体的，每一置信度均可以在一定程度上衡量训练文本的分类难度，为了减少单体决策导致的偏差，本步骤中可以综合考虑k-1个置信度，来确定训练文本的最终分类难度。

[0084] 一种可选的实施方式下，可以将各初始文本分类模型预测的所述训练文本属于标注类别的置信度的集合，确定为所述训练文本的分类难度表示。

[0085] 也即，可以将k-1个置信度的集合，确定为训练文本的分类难度表示。

[0086] 另一种可选的实施方式下，可以对各初始文本分类模型预测的所述训练文本属于标注类别的置信度进行数学运算，得到综合置信度，并将所述综合置信度确定为所述训练文本的分类难度表示。

[0087] 其中，对各个置信度进行数学运算可以包含多种不同的方式，如求取多个置信度的平均值、中位值、最大值、最小值等，将数学运算结果作为综合置信度，并将该综合置信度确定为训练文本的分类难度表示。

[0088] 当然，上述仅仅示例了两种可选的确定训练文本的分类难度的方式，除此之外，还可以采用其它可选的方式，以实现基于各初始文本分类模型预测的所述训练文本属于标注类别的置信度，确定所述训练文本的分类难度的目的。

[0089] 在本申请的一些实施例中，进一步对上述步骤S120，按照分类难度，将所述训练文本集中各训练文本重新划分为若干份不同分类难度的训练文本子集的过程进行介绍。

[0090] 本申请实施例中介绍了上述步骤S120的两种可选的实现方式，分别如下：

[0091] 第一种、

[0092] S1、参考设定的各分类难度区间，确定所述训练文本集中各训练文本的分类难度所处的分类难度区间。

[0093] 具体的，本申请实施例中可以预先设定多个不同的分类难度区间，各分类难度区间的区间范围可以相同，也可以不同。

[0094] 在划分好分类难度区间之后，针对训练文本集中各训练文本，可以依据训练文本的分类难度确定其所处的分类难度区间。

[0095] S2、将所述训练文本集中每一训练文本划分至所处分类难度区间对应的新训练文本子集，以得到每一分类难度区间对应的新训练文本子集。

[0096] 具体的，训练文本集中每一训练文本按照其分类难度所处的分类难度区间，可以将训练文本划分至分类难度区间对应的新训练文本子集中。也即，对应于每一分类难度区间，设置有一新训练文本子集。

[0097] 示例如：

[0098] 预先划定的分类难度区间有N个，对应于每一分类难度区间设置有一新训练文本子集。则针对训练文本集中每一训练文本，根据其分类难度所处的分类难度区间，可以将其划分至对应的新训练文本子集中，最终可以得到N个新训练文本子集C_i(i＝1,2,...,N)。其

中，每个新训练文本子集中包含训练文本的数量表示为Num(C_i)。

[0099] 一种可选的情况下，N个分类难度区间可以是：[0,0.2)，[0.2,0.4)，[0.4,0.5)，[0.5,0.6)，[0.6,0.7)，[0.7,0.8)，[0.8,0.9)，[0.9,0.95)，[0.95,1.0)。

[0100] 上述在确定训练文本的分类难度所属的分类难度区间时，可以根据训练文本的分类难度的不同表示形式，设计对应的方案。

[0101] 以前述实施例中介绍的两种不同的训练文本的分类难度表示形式为例进行说明：

[0102] 当训练文本的分类难度表示为多个置信度的集合时，则可以采用投票原则，决定训练文本的分类难度所属的分类难度区间。

[0103] 具体的，以置信度集合中每一置信度作为投票对象，确定每一投票对象所属的分类难度区间。对于每一分类难度区间，每当有一个投票对象落入当前分类难度区间时，当前分类难度区间的票数加1。当所有投票对象投票结束后，统计各个分类难度区间的最终票数，并选取最终票数最高的分类难度区间，作为训练文本的分类难度所属的分类难度区间。

[0104] 需要说明的是，当存在多个分类难度区间的最终票数相同时，为了一致性考虑，可以统一选择分类难度较小的一个分类难度区间作为训练文本的分类难度所属的分类难度区间。或者是，可以统一选择分类难度较大的一个分类难度区间作为训练文本的分类难度所属的分类难度区间。

[0105] 第二种、

[0106] S1、以分类难度作为聚类条件，对所述训练文本集中各训练文本进行聚类，得到聚类后的若干聚类簇。

[0107] 具体的，通过使用分类难度这一属性作为聚类条件对各训练文本进行聚类，聚类后的每一聚类簇中包含的是分类难度相同或相近的训练文本，而不同聚类簇中训练文本的分类难度差异较大。

[0108] S2、将每一聚类簇中的训练文本作为一个新训练文本子集，并基于聚类簇内各训练文本的分类难度确定聚类簇对应的新训练文本子集的分类难度。

[0109] 具体的，针对每一新训练文本子集，可以将其中包含的各训练文本的最大分类难度及最小分类难度组成的分类难度区间，作为新训练文本子集对应的分类难度。或者是，可以确定聚类中心对应的分类难度，作为新训练文本子集的分类难度。再或者是，可以将其中包含的各训练文本的分类难度求取平均值、中位值、最大值、最小值等，结果作为新训练文本子集对应的分类难度。

[0110] 示例如，通过聚类最终可以得到N个新训练文本子集C_i(i＝1,2,...,N)。其中，每个新训练文本子集中包含训练文本的数量表示为Num(C_i)。

[0111] 在本申请的另一实施例中，进一步提供了一种文本分类模型的训练方法。本实施例中的文本分类模型的训练方法，所使用的训练数据为前述实施例介绍的文本分类方法所划分后的各不同分类难度的训练文本子集。

[0112] 由于不同训练文本子集的分类难度不同，因此可以按照分类难度由低至高的顺序，依次利用各训练文本子集训练目标文本分类模型。

[0113] 本实施例的文本分类模型的训练方法，通过基于不同分类难度的训练文本子集，按照分类难度由低至高顺序递进式训练目标文本分类模型，这种渐进式的学习本质上模拟人脑学习新知识的过程，即由易到难、循序渐进，从而解决由于训练文本难度不均衡现象导

致的模型训练效果不佳的问题。

[0114] 可选的，上述按照分类难度由低至高的顺序，依次利用各训练文本子集训练目标文本分类模型的过程，可以包括：

[0115] 将目标文本分类模型的训练过程划分为若干阶段，每一阶段使用一个训练文本子集进行训练，且各训练阶段所使用的训练文本子集按照分类难度由低至高的顺序。

[0116] 另一种可选的实施方式，上述按照分类难度由低至高的顺序，依次利用各训练文本子集训练目标文本分类模型的过程，可以包括：

[0117] 按照分类难度由低至高的顺序，采用增量训练方式依次利用各训练文本子集训练目标文本分类模型。

[0118] 具体的，可以按照各训练文本子集的分类难度由低至高顺序，确定与每一训练文本子集对应的训练阶段。定义训练文本子集有N个，表示为C_i(i＝1,2,...,N)。则训练阶段也有N个，分别表示为:S_i(i＝1,2,...,N)。

[0119] 在每一训练阶段中，从所述训练阶段及之前各训练阶段分别对应的训练文本子集中采样训练文本，采样结果组成所述训练阶段对应的训练文本集合。

[0120] 假设当前训练阶段为S_i，则从C_1,C_2,...C_i中分别采样训练文本，采样结果组成当前训练阶段S_i对应的训练文本集合。

[0121] 其中可选的，从C_1,C_2,...C_i中每个训练文本子集中采样的训练文本的个数可以相同或不同，示例如，从C_1,C_2,...C_i中每个训练文本子集中分别采样$\frac{1}{N}num(C_i)$个训练文本，由采样结果组成当前训练阶段S_i对应的训练文本集合。

[0122] 在确定了每一训练阶段对应的训练文本集合之后，可以利用每一训练阶段对应的训练文本集合，依次训练目标文本分类模型。

[0123] 本实施例中提供的模型训练方法，通过采用增量式训练方式对模型进行训练，避免模型出现灾难性遗忘，导致模型忘记之前训练阶段所学到的知识。

[0124] 进一步可选的，在依次利用各训练文本子集训练目标文本分类模型之后，本申请方案还可以进一步增加全量数据训练过程，也即：

[0125] 以各不同难度的训练文本子集组成全量训练文本集，并利用所述全量训练文本集训练目标文本分类模型。

[0126] 通过进一步增加使用全量训练文本集对目标文本分类模型进行训练，能够使得模型收敛性更好，也即使得目标文本分类模型的性能更佳。

[0127] 在本申请的一些实施例中，对上述每一训练阶段，使用训练文本集合训练目标文本分类模型的过程进行介绍。

[0128] 一种直接的方式如，直接利用训练文本集合对目标文本分类模型进行训练。

[0129] 此外，考虑到现实场景中，大量的训练文本还可能存在样本分类不均衡问题，也即大多数分类标签下只含有少量的训练文本数据。这种由于不同类别的标注数据缺失，会导致基于此训练出的目标文本分类模型，容易受到头部标签数据的不良偏见影响，影响目标文本分类模型的分类结果准确度。

[0130] 为此，本实施例中提供了一种解决方案，即，在每一训练阶段中，利用训练阶段对应的训练文本集合，以及训练文本集合中各类别文本的分布信息，训练目标文本分类模型。

[0131] 相比于前一种方式，后一种训练方式在目标文本分类模型的训练过程中，融入了训练文本集合中各类别文本的先验分布信息，也即将训练文本集合中各类别文本的先验分布信息融入建模过程中，可以进一步缓解不同标签类别分布不均衡造成的问题，可以降低目标文本分类模型对头部标签的过拟合现象。

[0132] 其中，上述将训练文本集合中各类别文本的先验分布信息融入建模过程，具体可以采用将各类别文本的先验分布信息以互信息的形式融入建模过程。

[0133] 其中，互信息表示两个随机变量之间的相关程度，其数学公式可以表示为：

[0134] $$\log\frac{p(x,y)}{p(x)p(y)}$$ 公式1

[0135] 对于目标文本分类模型，其常用的分类损失函数都是交叉熵损失函数，其可以表示为：

[0136] $$p(y|x)=\frac{e^{f(y|x)}}{\sum_{m} e^{f_m(y|x)}}$$ 公式2

[0137] 其中，p(y|x)表示给定输入样本x，预测类别为y的条件概率，m为分类标签个数，f(y|x)表示给定输入样本x，模型在类别y上的预测值。

[0138] 类比于互信息，本申请可以得到如下公式：

[0139] $$\log\frac{p(y|x)}{p(y)}\propto f_y(x,\Theta)\Leftrightarrow \log p_\Theta(y|x)=f_y(x,\Theta)+\log(p(y))$$ 公式3

[0140] 其中，$f_y(x,\Theta)$表示给定输样本入x，模型在每个类别y上的预测值，log(p(y))表示类别y的先验分布信息。

[0141] 因此，目标文本分类模型在对每个输入训练文本预测的得分进行softmax归一化之前，可以进一步加上每个类别的先验分布信息，从而缓解因为类别分布不均给模型带来的影响，将类别分布的先验分布信息加入到模型中，可以让模型聚焦于先验分布信息所解决不了的本质部分。

[0142] 目标文本分类模型的得分是一个概率值，这是对模型输出进行归一化的过程，而模型学习的本质是对p(yx)的建模过程，大多数神经网络都是对条件概率进行建模，但是相比于拟合条件概率，如果能让模型直接拟合互信息，那么模型将会学到更本质的知识，但是相比于拟合条件概率(直接用交叉熵进行训练即可)，拟合互信息不容易进行训练。因此，本申请实施例提供了一种较优的方案，即使目标文本分类模型任然使用交叉熵为损失函数，但本质是在拟合互信息。具体的，上述公式1-3的推导过程就是在建模互信息和条件概率之间的关系，通过这种方式可以将训练文本集合中各类别文本先验分布信息融入模型中，使得模型在训练的时候可以关注到其它知识，缓解样本类别不均衡的问题。

[0143] 对于上述公式3，将公式3按照右端的形式进行归一化处理，可以得到：

[0144] $$p_\theta(y\,|\,x)=\frac{e^{f_y(x;\theta)+\log p(y)}}{\sum_{i=1}^{m} e^{f_i(x;\theta)+\log(p(i))}}=\log[1+\sum_{i\neq y}\left(\frac{p(i)}{p(y)}\right)^{\tau} e^{f_i(x;\theta)-f_y(x;\theta)}]$$ 公式4

[0145] 因此，可以使用$p_\theta(y|x)$作为目标文本分类模型最终的损失函数loss，也即loss表

示为：

[0146] $$loss = \log[1 + \sum_{i \neq y} \left(\frac{p(i)}{p(y)} \right)^{\tau} e^{f_i(x;\theta) - f_y(x;\theta)}]$$ 公式5

[0147] 目标文本分类模型的训练过程可以结合图2进行说明，则利用所述训练阶段对应的训练文本集合，以及所述训练文本集合中各类别文本的分布信息，训练目标文本分类模型的过程，可以包括如下步骤：

[0148] S1、将所述训练阶段对应的训练文本集合中每一训练文本的编码特征输入目标文本分类模型。

[0149] S2、由所述目标文本分类模型的隐层对所述编码特征进行处理，并基于处理后的编码特征预测所述训练文本在各分类标签上的得分。

[0150] 其中，目标文本分类模型的隐层可以包括一个BERT编码器，以及若干个全连接层。

[0151] 隐层对训练文本的编码特征进行特征处理，得到处理后的编码特征，并基于该处理后的编码特征来预测训练文本在各个分类标签上的得分。图2中示例的是共有m个分类标签。

[0152] 隐层中最后一个全连接层会输出训练文本在1-m个分类标签上各自的得分，该得分是未经过归一化的概率得分logtis。

[0153] S3、由所述目标文本分类模型将所述训练文本在每一分类标签上的得分，与所述训练文本集合中对应分类标签下训练文本的数量相加，得到所述训练文本在每一分类标签上的综合得分，并对各分类标签上的综合得分进行归一化处理。

[0154] 具体的，训练文本集合中各类别的先验分布信息也即，训练文本集合中每一分类标签下训练文本的数量，对应图2中类别先验分布信息log(p(y))。

[0155] 本步骤中，对于训练文本在0-m个分类标签上未归一化的得分，分别与对应分类标签下训练文本的数量相加，结果作为训练文本在对应分类标签上的综合得分，也即最终的finalscore。

[0156] S4、结合所述训练文本在各分类标签上归一化后的综合得分，以所述训练文本的分类损失为损失函数，训练目标文本分类模型。

[0157] 具体的，目标文本分类模型的损失函数可以使用上述公式5所示的，融合了训练文本集合中各类别文本的分布信息的互信息的损失函数。

[0158] 下面对本申请实施例提供的文本分类装置进行描述，下文描述的文本分类装置与上文描述的文本分类方法可相互对应参照。

[0159] 参见图3，图3为本申请实施例公开的一种文本分类装置结构示意图。

[0160] 如图3所示，该装置可以包括：

[0161] 初始文本分类模型训练单元11，用于将训练文本集划分为若干份训练文本子集，并利用每一份训练文本子集训练对应的初始文本分类模型；

[0162] 分类难度预测单元12，用于分别以每一份训练文本子集为验证集，利用除所述验证集外的其它各训练文本子集对应的初始文本分类模型，对所述验证集中每一训练文本进行分类预测，得到验证集中每一训练文本的分类难度；

[0163] 训练文本再划分单元13，用于按照分类难度，将所述训练文本集中各训练文本重新划分为若干份不同分类难度的训练文本子集。

[0164] 可选的，上述初始文本分类模型训练单元将训练文本集划分为若干份训练文本子集的过程，可以包括：

[0165] 采用随机划分的方式，将训练文本集随机划分为若干份互不交叉的训练文本子集。

[0166] 可选的，上述分类难度预测单元利用各初始文本分类模型，对验证集中每一训练文本进行分类预测，得到每一训练文本的分类难度的过程，可以包括：

[0167] 针对验证集中每一训练文本，利用各初始文本分类模型，分别对所述训练文本进行分类预测，得到每一初始文本分类模型预测的所述训练文本属于标注类别的置信度；

[0168] 基于各初始文本分类模型预测的所述训练文本属于标注类别的置信度，确定所述训练文本的分类难度。

[0169] 可选的，上述分类难度预测单元基于各初始文本分类模型预测的所述训练文本属于标注类别的置信度，确定所述训练文本的分类难度的过程，可以包括：

[0170] 将各初始文本分类模型预测的所述训练文本属于标注类别的置信度的集合，确定为所述训练文本的分类难度表示；

[0171] 或，

[0172] 对各初始文本分类模型预测的所述训练文本属于标注类别的置信度进行数学运算，得到综合置信度，并将所述综合置信度确定为所述训练文本的分类难度表示。

[0173] 可选的，本申请实施例公开了上述训练文本再划分单元按照分类难度，将所述训练文本集中各训练文本重新划分为若干份不同分类难度的训练文本子集的两种不同实施方式，分别如下：

[0174] 第一种实施方式可以包括：

[0175] 参考设定的各分类难度区间，确定所述训练文本集中各训练文本的分类难度所处的分类难度区间；

[0176] 将所述训练文本集中每一训练文本划分至所处分类难度区间对应的新训练文本子集，以得到每一分类难度区间对应的新训练文本子集。

[0177] 第二种实施方式可以包括：

[0178] 以分类难度作为聚类条件，对所述训练文本集中各训练文本进行聚类，得到聚类后的若干聚类簇；

[0179] 将每一聚类簇中的训练文本作为一个新训练文本子集，并基于聚类簇内各训练文本的分类难度确定聚类簇对应的新训练文本子集的分类难度。

[0180] 下面对本申请实施例提供的文本分类模型训练装置进行描述，下文描述的文本分类模型训练装置与上文描述的文本分类模型训练方法可相互对应参照。

[0181] 参见图4，图4为本申请实施例公开的一种文本分类模型训练装置结构示意图。

[0182] 如图4所示，该装置可以包括：

[0183] 训练文本子集获取单元21，用于获取前述实施例的文本分类方法所划分的各不同难度的训练文本子集；

[0184] 目标文本分类模型训练单元22，用于按照分类难度由低至高的顺序，依次利用各训练文本子集训练目标文本分类模型。

[0185] 可选的，上述目标文本分类模型训练单元按照分类难度由低至高的顺序，依次利

用各训练文本子集训练目标文本分类模型的过程，可以包括：

[0186] 按照分类难度由低至高的顺序，采用增量训练方式依次利用各训练文本子集训练目标文本分类模型。

[0187] 可选的，本申请的文本分类模型训练装置还可以包括：

[0188] 全量训练单元，用于在所述依次利用各训练文本子集训练目标文本分类模型之后，以各不同难度的训练文本子集组成全量训练文本集，并利用所述全量训练文本集训练所述目标文本分类模型。

[0189] 可选的，上述目标文本分类模型训练单元按照分类难度由低至高的顺序，采用增量训练方式依次利用各训练文本子集训练目标文本分类模型的过程，可以包括：

[0190] 按照各训练文本子集的分类难度由低至高顺序，确定与每一训练文本子集对应的训练阶段；

[0191] 在每一训练阶段中，从所述训练阶段及之前各训练阶段分别对应的训练文本子集中采样训练文本，采样结果组成所述训练阶段对应的训练文本集合；

[0192] 利用所述训练阶段对应的训练文本集合，训练目标文本分类模型。

[0193] 可选的，上述目标文本分类模型训练单元利用所述训练阶段对应的训练文本集合，训练目标文本分类模型的过程，可以包括：

[0194] 利用所述训练阶段对应的训练文本集合，以及所述训练文本集合中各类别文本的分布信息，训练目标文本分类模型。

[0195] 可选的，上述目标文本分类模型利用所述训练阶段对应的训练文本集合，以及所述训练文本集合中各类别文本的分布信息，训练目标文本分类模型的过程，可以包括：

[0196] 将所述训练阶段对应的训练文本集合中每一训练文本的编码特征输入目标文本分类模型；

[0197] 由所述目标文本分类模型的隐层对所述编码特征进行处理，并基于处理后的编码特征预测所述训练文本在各分类标签上的得分；

[0198] 由所述目标文本分类模型将所述训练文本在每一分类标签上的得分，与所述训练文本集合中对应分类标签下训练文本的数量相加，得到所述训练文本在每一分类标签上的综合得分，并对各分类标签上的综合得分进行归一化处理；

[0199] 结合所述训练文本在各分类标签上归一化后的综合得分，以所述训练文本的分类损失为损失函数，训练目标文本分类模型。

[0200] 本申请实施例提供的文本分类装置可应用于文本分类设备，如终端：手机、电脑等。可选的，图5示出了文本分类设备的硬件结构框图，参照图5，文本分类设备的硬件结构可以包括：至少一个处理器1，至少一个通信接口2，至少一个存储器3和至少一个通信总线4；

[0201] 在本申请实施例中，处理器1、通信接口2、存储器3、通信总线4的数量为至少一个，且处理器1、通信接口2、存储器3通过通信总线4完成相互间的通信；

[0202] 处理器1可能是一个中央处理器CPU，或者是特定集成电路ASIC(Application Specific Integrated Circuit)，或者是被配置成实施本发明实施例的一个或多个集成电路等；

[0203] 存储器3可能包含高速RAM存储器，也可能还包括非易失性存储器(non-volatile

memory)等,例如至少一个磁盘存储器;

[0204] 其中,存储器存储有程序,处理器可调用存储器存储的程序,所述程序用于:

[0205] 将训练文本集划分为若干份训练文本子集,并利用每一份训练文本子集训练对应的初始文本分类模型;

[0206] 分别以每一份训练文本子集为验证集,利用除所述验证集外的其它各训练文本子集对应的初始文本分类模型,对所述验证集中每一训练文本进行分类预测,得到验证集中每一训练文本的分类难度;

[0207] 按照分类难度,将所述训练文本集中各训练文本重新划分为若干份不同分类难度的训练文本子集。

[0208] 可选的,所述程序的细化功能和扩展功能可参照上文描述。

[0209] 本申请实施例还提供一种存储介质,该存储介质可存储有适于处理器执行的程序,所述程序用于:

[0210] 将训练文本集划分为若干份训练文本子集,并利用每一份训练文本子集训练对应的初始文本分类模型;

[0211] 分别以每一份训练文本子集为验证集,利用除所述验证集外的其它各训练文本子集对应的初始文本分类模型,对所述验证集中每一训练文本进行分类预测,得到验证集中每一训练文本的分类难度;

[0212] 按照分类难度,将所述训练文本集中各训练文本重新划分为若干份不同分类难度的训练文本子集。

[0213] 可选的,所述程序的细化功能和扩展功能可参照上文描述。

[0214] 最后,还需要说明的是,在本文中,诸如第一和第二等之类的关系术语仅仅用来将一个实体或者操作与另一个实体或操作区分开来,而不一定要求或者暗示这些实体或操作之间存在任何这种实际的关系或者顺序。而且,术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含,从而使得包括一系列要素的过程、方法、物品或者设备不仅包括那些要素,而且还包括没有明确列出的其他要素,或者是还包括为这种过程、方法、物品或者设备所固有的要素。在没有更多限制的情况下,由语句“包括一个……”限定的要素,并不排除在包括所述要素的过程、方法、物品或者设备中还存在另外的相同要素。

[0215] 本说明书中各个实施例采用递进的方式描述,每个实施例重点说明的都是与其他实施例的不同之处,各个实施例之间可以根据需要进行组合,且相同相似部分互相参见即可。

[0216] 对所公开的实施例的上述说明,使本领域专业技术人员能够实现或使用本申请。对这些实施例的多种修改对本领域的专业技术人员来说将是显而易见的,本文中所定义的一般原理可以在不脱离本申请的精神或范围的情况下,在其它实施例中实现。因此,本申请将不会被限制于本文所示的这些实施例,而是要符合与本文所公开的原理和新颖特点相一致的最宽的范围。

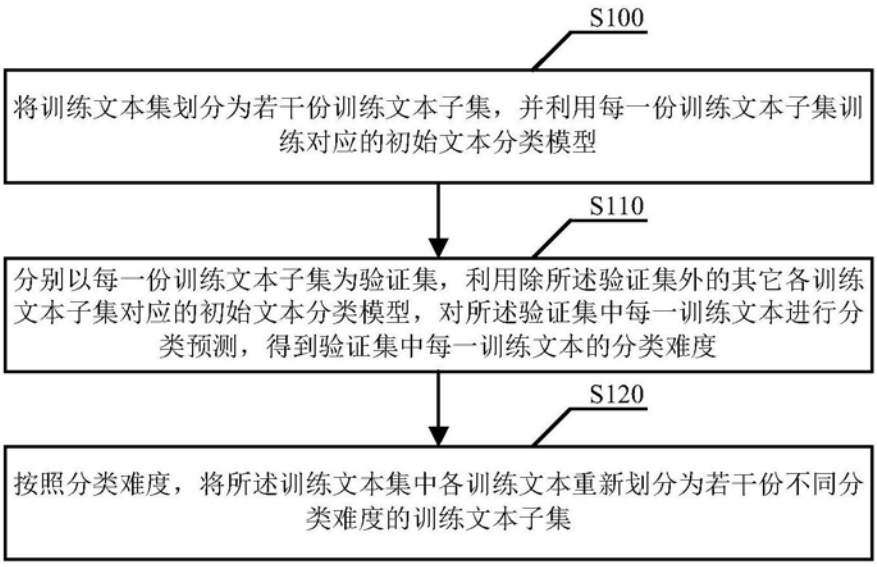

S100
将训练文本集划分为若干份训练文本子集，并利用每一份训练文本子集训练对应的初始文本分类模型
S110
分别以每一份训练文本子集为验证集，利用除所述验证集外的其它各训练文本子集对应的初始文本分类模型，对所述验证集中每一训练文本进行分类预测，得到验证集中每一训练文本的分类难度
S120
按照分类难度，将所述训练文本集中各训练文本重新划分为若干份不同分类难度的训练文本子集

图1

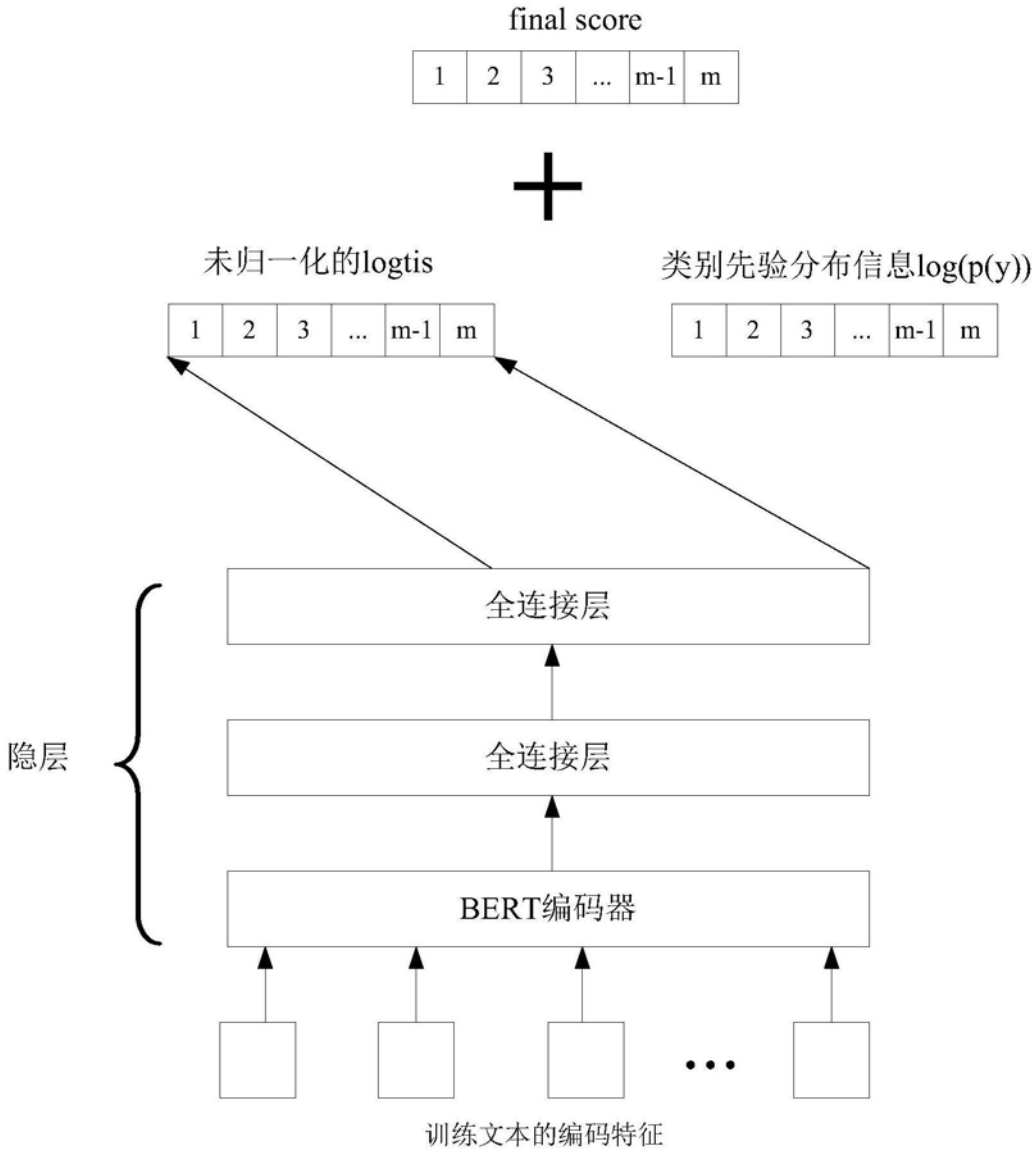

final score
1 2 3 ... m-1 m
未归一化的logtis
1 2 3 ... m-1 m
类别先验分布信息log(p(y))
1 2 3 ... m-1 m
全连接层
全连接层
隐层
BERT编码器
训练文本的编码特征

图2

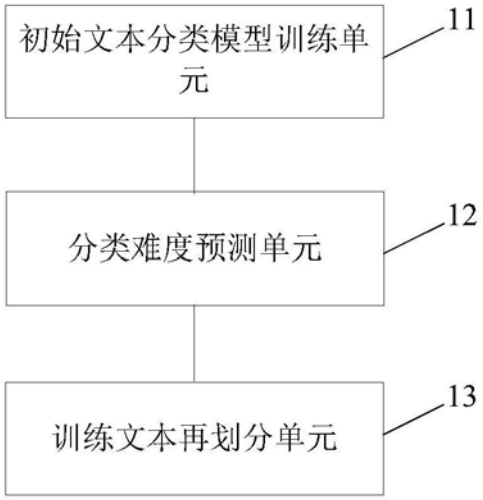

初始文本分类模型训练单元
11
分类难度预测单元
12
训练文本再划分单元
13

图3

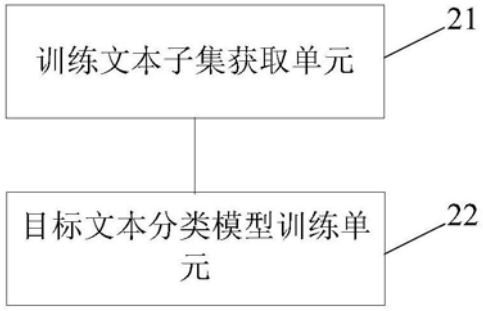

训练文本子集获取单元
21
目标文本分类模型训练单元
22

图4

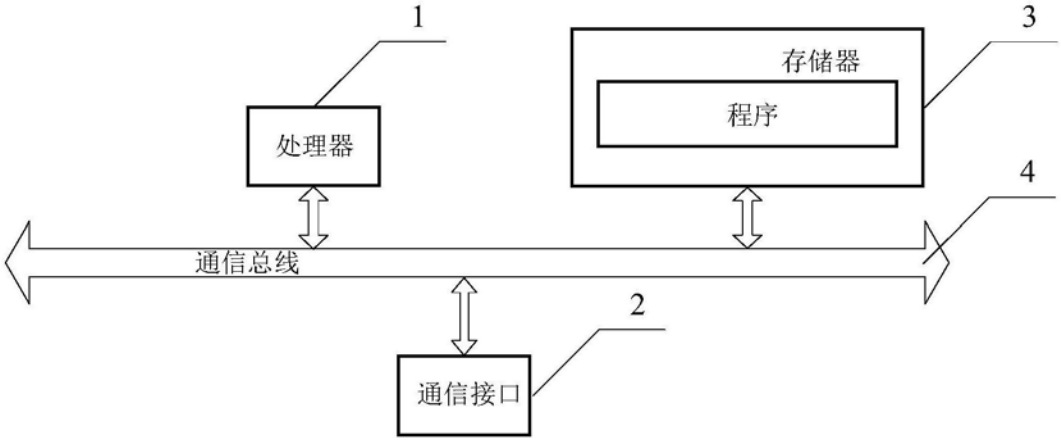

1
处理器
3
存储器
程序
4
通信总线
2
通信接口

图5