



(12)发明专利申请

(10)申请公布号 CN 111475622 A

(43)申请公布日 2020.07.31

(21)申请号 202010268806.4

(22)申请日 2020.04.08

(71)申请人 广东工业大学

地址 510060 广东省广州市越秀区东风东路729号大院

(72)发明人 王涛 周佳乐 邓健峰

(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 沈闯

(51)Int.Cl.

G06F 16/33(2019.01)

G06F 16/35(2019.01)

G06F 40/289(2020.01)

G06F 40/30(2020.01)

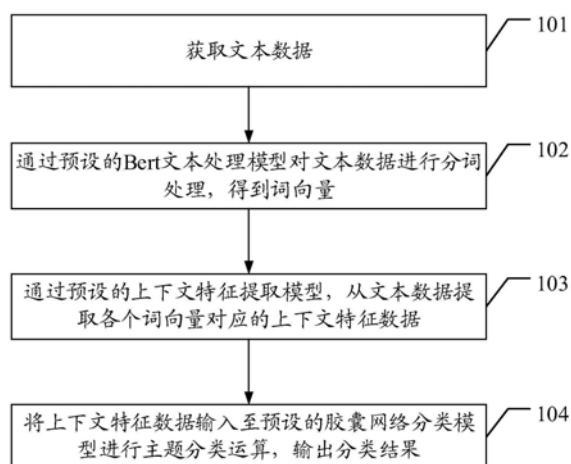
权利要求书2页 说明书7页 附图4页

(54)发明名称

一种文本分类方法、装置、终端及存储介质

(57)摘要

本申请提供了一种文本分类方法、装置、终端及存储介质,其中方法包括:获取文本数据;通过预设的Bert文本处理模型对所述文本数据进行分词处理,得到词向量;通过预设的上下文特征提取模型,从所述文本数据提取各个所述词向量对应的上下文特征数据;将所述上下文特征数据输入至预设的胶囊网络分类模型进行主题分类运算,输出分类结果。本申请采用结合Bert、GAT、胶囊网络的方法,通过Bert预处理,GAT学习文本上下文特征,然后通过胶囊网络对特征进行主题分类通过胶囊网络中动态路由的过程,减少了卷积神经网络中数据在反向传播过程中可能给文本分类器模型带来的一些弊端,从而提高文本分类的准确性,解决了现有的深度学习方法文本分类准确率低的技术问题。



1. 一种文本分类方法,其特征在于,包括:
获取文本数据;
通过预设的Bert文本处理模型对所述文本数据进行分词处理,得到词向量;
通过预设的上下文特征提取模型,从所述文本数据提取各个所述词向量对应的上下文特征数据;
将所述上下文特征数据输入至预设的胶囊网络分类模型进行主题分类运算,输出分类结果。
2. 根据权利要求1所述的一种文本分类方法,其特征在于,所述获取文本数据之后还包括:
对所述文本数据进行预处理,其中所述预处理包括,去除停用词和去除标点符号。
3. 根据权利要求1所述的一种文本分类方法,其特征在于,所述上下文特征提取模型具体为GAT图注意网络模型。
4. 根据权利要求1所述的一种文本分类方法,其特征在于,还包括:
将预置的训练样本数据输入至初始胶囊网络模型,对所述初始胶囊网络模型进行训练,得到所述胶囊网络分类模型,其中,所述训练样本数据为将预置的样本文本数据通过所述Bert文本处理模型和所述上下文特征提取模型处理得到的上下文特征样本数据。
5. 一种文本分类装置,其特征在于,包括:
文本获取单元,用于获取文本数据;
分词处理单元,用于通过预设的Bert文本处理模型对所述文本数据进行分词处理,得到词向量;
上下文特征提取单元,用于通过预设的上下文特征提取模型,从所述文本数据提取各个所述词向量对应的上下文特征数据;
文本分类单元,用于将所述上下文特征数据输入至预设的胶囊网络分类模型进行主题分类运算,输出分类结果。
6. 根据权利要求5所述的一种文本分类装置,其特征在于,所述获取文本数据之后还包括:
预处理单元,用于对所述文本数据进行预处理,其中所述预处理包括,去除停用词和去除标点符号。
7. 根据权利要求5所述的一种文本分类装置,其特征在于,所述上下文特征提取模型具体为GAT图注意网络模型。
8. 根据权利要求5所述的一种文本分类装置,其特征在于,还包括:
胶囊网络分类模型训练单元,用于将预置的训练样本数据输入至初始胶囊网络模型,对所述初始胶囊网络模型进行训练,得到所述胶囊网络分类模型,其中,所述训练样本数据为将预置的样本文本数据通过所述Bert文本处理模型和所述上下文特征提取模型处理得到的上下文特征样本数据。
9. 一种终端,其特征在于,包括:存储器和处理器;
所述存储器用于存储与权利要求1至4任意一项所述的文本分类方法对应的程序代码;
所述处理器用于执行所述程序代码。
10. 一种存储介质,其特征在于,所述存储介质中保存有与权利要求1至4任意一项所述

的文本分类方法对应的程序代码。

一种文本分类方法、装置、终端及存储介质

技术领域

[0001] 本申请涉及文本分类领域,尤其涉及一种文本分类方法、装置、终端及存储介质。

背景技术

[0002] 伴随着大数据时代的发展,产生积累了大量的文本数据,文本数据涉及到的类别主题各式各样,通用领域当中,通过对这些文本数据分析处理,可以更深入的了解其中的关系。然而,在得到的各类文本中多种多样,难以被计算机理解,涉及到的主题有所不同,假设能把这些文本进行有效的分类识别,对大数据处理工作有很大的正向作用。

[0003] 目前,用于文本分类的主流的方法有:GNN、CNN。循环神经网络GNN虽然能够获取文本的上下文特征,但它存在梯度爆炸的问题;卷积神经网络CNN则存在池化层中会丢失信息,无法获取文本全局信息等问题,导致现有的深度学习方法难以获得较高的文本分类准确率。

发明内容

[0004] 本申请提供了一种文本分类方法、装置、终端及存储介质,用于解决现有的深度学习方法文本分类准确率低的技术问题。

[0005] 有鉴于此,本申请第一方面提供了一种文本分类方法,包括:

[0006] 获取文本数据;

[0007] 通过预设的Bert文本处理模型对所述文本数据进行分词处理,得到词向量;

[0008] 通过预设的上下文特征提取模型,从所述文本数据提取各个所述词向量对应的上下文特征数据;

[0009] 将所述上下文特征数据输入至预设的胶囊网络分类模型进行主题分类运算,输出分类结果。

[0010] 可选地,所述获取文本数据之后还包括:

[0011] 对所述文本数据进行预处理,其中所述预处理包括,去除停用词和去除标点符号。

[0012] 可选地,所述上下文特征提取模型具体为GAT图注意网络模型。

[0013] 可选地,还包括:

[0014] 将预置的训练样本数据输入至初始胶囊网络模型,对所述初始胶囊网络模型进行训练,得到所述胶囊网络分类模型,其中,所述训练样本数据为将预置的样本文本数据通过所述Bert文本处理模型和所述上下文特征提取模型处理得到的上下文特征样本数据。

[0015] 本申请第二方面提供了一种文本分类装置,包括:

[0016] 文本获取单元,用于获取文本数据;

[0017] 分词处理单元,用于通过预设的Bert文本处理模型对所述文本数据进行分词处理,得到词向量;

[0018] 上下文特征提取单元,用于通过预设的上下文特征提取模型,从所述文本数据提取各个所述词向量对应的上下文特征数据;

[0019] 文本分类单元,用于将所述上下文特征数据输入至预设的胶囊网络分类模型进行主题分类运算,输出分类结果。

[0020] 可选地,所述获取文本数据之后还包括:

[0021] 预处理单元,用于对所述文本数据进行预处理,其中所述预处理包括,去除停用词和去除标点符号。

[0022] 可选地,所述上下文特征提取模型具体为GAT图注意网络模型。

[0023] 可选地,还包括:

[0024] 胶囊网络分类模型训练单元,用于将预置的训练样本数据输入至初始胶囊网络模型,对所述初始胶囊网络模型进行训练,得到所述胶囊网络分类模型,其中,所述训练样本数据为将预置的样本文本数据通过所述Bert文本处理模型和所述上下文特征提取模型处理得到的上下文特征样本数据。

[0025] 本申请第三方面提供了一种终端,包括:存储器和处理器;

[0026] 所述存储器用于存储与本申请第一方面所述的文本分类方法对应的程序代码;

[0027] 所述处理器用于执行所述程序代码。

[0028] 本申请第四方面提供了一种存储介质,所述存储介质中保存有与本申请第一方面所述的文本分类方法对应的程序代码。

[0029] 从以上技术方案可以看出,本申请实施例具有以下优点:

[0030] 本申请提供了一种文本分类方法,包括:获取文本数据;通过预设的Bert文本处理模型对所述文本数据进行分词处理,得到词向量;通过预设的上下文特征提取模型,从所述文本数据提取各个所述词向量对应的上下文特征数据;将所述上下文特征数据输入至预设的胶囊网络分类模型进行主题分类运算,输出分类结果。

[0031] 本申请采用结合Bert、GAT、胶囊网络的方法,通过Bert预处理,GAT学习文本上下文特征,然后通过胶囊网络对特征进行主题分类通过胶囊网络中动态路由的过程,减少了卷积神经网络中数据在反向传播过程中可能给文本分类器模型带来的一些弊端,从而提高文本分类的准确性,解决了现有的深度学习方法文本分类准确率低的技术问题。

附图说明

[0032] 为了更清楚地说明本申请实施例或现有技术中的技术方案,下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍,显而易见地,下面描述中的附图仅仅是本申请的一些实施例,对于本领域普通技术人员来讲,在不付出创造性劳动性的前提下,还可以根据这些附图获得其它的附图。

[0033] 图1为本申请提供的一种文本分类方法的第一个实施例的流程示意图;

[0034] 图2为本申请提供的一种文本分类方法的第二个实施例的流程示意图;

[0035] 图3为本申请提供的一种文本分类装置的第一个实施例的结构示意图;

[0036] 图4为GAT提取文本上下文特征流程图;

[0037] 图5为胶囊网络主题分类示例图。

具体实施方式

[0038] 本申请实施例提供了一种文本分类方法、装置、终端及存储介质,用于解决现有的

深度学习方法文本分类准确率低的技术问题。

[0039] 为使得本申请的发明目的、特征、优点能够更加的明显和易懂,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,下面所描述的实施例仅仅是本申请一部分实施例,而非全部的实施例。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它实施例,都属于本申请保护的范围。

[0040] 请参阅图1,本申请第一方面提供了一种文本分类方法,包括:

[0041] 步骤101、获取文本数据。

[0042] 步骤102、通过预设的Bert文本处理模型对文本数据进行分词处理,得到词向量。

[0043] 需要说明的是,本实施例在获取到待分类处理的文本数据后,首先利用Bert文本处理模型对文本数据进行第一步分词处理,包括词嵌入,将输入文本转为向量输出,得到词向量。

[0044] Bert模型首先对input进行编码,转为模型需要的编码格式,使用辅助标记符[CLS]和[SEP]来表示句子的开始和分隔。然后根据输入得到对应的embedding,每个单词有三个embedding:位置信息embedding,单词embedding,句子embedding,把单词对应的三个embedding叠加,就形成了GAT的输入。

[0045] 其中,本实施例采用的Bert文本处理模型是一个句子级别的语言模型,采用Bert文本处理模型可以进一步增加词向量模型泛化能力,充分描述字符级、词级、句子级甚至句间关系特征,大大提高了模型表示的效率。

[0046] 步骤103、通过预设的上下文特征提取模型,从文本数据提取各个词向量对应的上下文特征数据。

[0047] 需要说明的是,如图4所示,本实施例采用的上下文特征提取模型具体为GAT图神经网络模型,GAT通过注意力机制对邻居节点做聚合操作,实现对不同邻居权重的分配,对重点内容配以更大的注意力权重,从而大大提高了图神经网络模型的表达能力。它的作用在于,把知识表述成图,并且用attention机制,对每个输入词向量进行attention计算,为每个节点分配不同权重,用权重参数体现邻域节点对中心节点的影响力大小,学习上下文句子之间的依赖关系,提取文本中的上下文特征,从而提高文本的语义表达能力,实现对文本的理解由浅层分析转向深度融合,发现文本中不同句子的注意力,可以快速过滤注意力较小的信息,减少计算任务的复杂度。关注作用比较大的节点,而忽视一些作用较小的节点。在处理局部信息的时候同时能够关注整体的信息,以便更加精确地对节点特征进行提取。

[0048] 更具体地,本实施例的利用隐藏的自注意层解决现有的基于图卷积及其类似方法的缺点。通过堆叠这样的一些层,层里的节点能够注意其邻近节点的特征,为邻近的不同节点指定不同的权重,将前一层得到的向量转化为图中的节点。主要由编码器和解码器两部分组成,编码器由若干对自注意力层组成,解码器结构同编码器大致一样,区别是在两层中间新添加了一个Attention层,用于聚焦当前解码单词对应的输入部分,主要是通过自注意力机制来表示句子的上下文特征,通过将得到的向量转化为节点表示,利用GAT的注意力层,GAT的计算也分为两步:计算注意力系数和加权求和。

[0049] 计算注意力系数:

[0050] 输入是一个节点特征向量集合 $h = \{\vec{h}_1, \vec{h}_2, \dots, \vec{h}_n\}$, $\vec{h}_i \in R^F$, N 代表节点数目, F 为每个

节点的特征数。该层产生一个新的节点特征集合 $h' = \{\vec{h}_1', \vec{h}_2', \dots, \vec{h}_n'\}$, $\vec{h}_i' \in R^{F'}$ 作为输出。为了

去获得足够的把输入的特征转化为更高层次的特征的表现力,至少需要一个可学习的线性转换。为此目的,作为一个初始化步骤,一个共享的参数为一个权重矩阵 $W \in R^{F' \times F}$ 的线性转换应用在每个节点上。然后在节点上使用自我注意地机制,一个共享的注意力机制 a , 以此来计算注意力因子

$$[0051] \quad e_{ij} = a(W \vec{h}_i', W \vec{h}_j')$$

[0052] 上面的公式表示节点 j 的特征对于节点 i 的重要性。在其最通用的表述中, j 将是 i 的第一阶邻居。为了使不同节点的因子容易比较,我们使用柔性最大值函数均一化

$$[0053] \quad a_{ij} = \text{soft max}_j(e_{ij}) = \frac{\exp(e_{ij})}{\sum_{k \in N_i} \exp(e_{ik})}$$

[0054] 一旦获得均一化的注意力因子,它将用来计算与之相关的特征的线性组合,结果作为每个节点的最终输出。

$$[0055] \quad \vec{h}_i' = \sigma\left(\sum_{j \in N_i} a_{ij} W \vec{h}_j'\right)$$

[0056] 加权求和:

[0057] 得到归一化的注意力系数后,使用归一化的值计算对应特征的线性组合,作为每个顶点最后的输出特征,将输出量输入到下一层网络中。这里巧妙的依赖到上下文的信息,使得任务与文本结合更加紧密,以提高准确率,表现出更好的特征抽取效果。

[0058] 例如输入一个句子,那么里面的每个词都要和该句子中的所有词进行attention计算。目的是学习句子内部的词依赖关系,捕获句子的内部结构。在长距离依赖上,由于self-attention是每个词和所有词都要计算attention,所以不管他们中间有多长距离,最大的路径长度也都只是1。可以捕获长距离依赖关系。

[0059] 步骤104、将上下文特征数据输入至预设的胶囊网络分类模型进行主题分类运算,输出分类结果。

[0060] 需要说明的是,如图5所示,本实施例的胶囊网络分类模型为利用胶囊网络的压缩函数,如式(1)所示,即每个胶囊的模长,而模长可以归一化到0和1之间,以此来表示分类的概率胶囊数量为 n 个。

$$[0061] \quad v_j = \frac{\|s_j\|^2}{1 + \|s_j\|^2} \frac{s_j}{\|s_j\|} \quad (1)$$

[0062] 其中, v_j 为胶囊最终输出向量, s_j ($j \in [0, m]$) 为低级胶囊的输出向量。基于上一层得到的文本向量进行表达,并作为胶囊网络的输入,训练胶囊网络文本分类模型具体为:

[0063] 对文本向量进行局部特征提取;

[0064] 对文本特征进行特征重构,从低维度空间将其映射到更高维度的空间中;

[0065] 利用胶囊层进行分类,胶囊层的输入神经元向量通过加权、耦合、挤压、动态路由在输入胶囊到输出胶囊之间流动;

[0066] 胶囊层的计算公式如(2)–(3)所示。

$$[0067] \quad s_j = \sum_i c_{ij} u_{j|i} \quad (2)$$

$$[0068] \quad u_{j|i} = w_{ij} l_i \quad (3)$$

[0069] 其中, w_{ij} 代表的是权重矩阵, $u_{j|i}$ 为低级胶囊的输出向量。 l_i ($i \in [0, N]$) 为胶囊网络的输入。 c_{ij} 为权重参数,可以通过胶囊网络中包含的动态路由算法计算得出,更具体的,本实施例的动态路由算法的输入为向量 $u_{j|i}$,迭代次数 r ,输出为分类胶囊向量 v_j 。

[0070] 首先,初始化。对所有低级胶囊 i 和高级胶囊 j ,设定参数 b_{ij} ,初始为0。胶囊层的输入神经元向量通过加权、耦合、挤压、动态路由在输入胶囊到输出胶囊之间流动。其次,针对每次迭代的过程,对每个低级胶囊 i 运算, $c_i = \text{softmax}(b_i)$,对每个高级胶囊 j ,有 $s_j = \sum_i c_{ij} u_{j|i}$,对高级胶囊输出进行挤压操作,有 $v_j = \text{squash}(s_j)$,最后更新 b_{ij} 的操作为 $b_{ij} = b_{ij} + u_{j|i} * v_j$ 。

[0071] 本发明针对传统网络在文本分类下存在梯度爆炸和无法获取文本全局信息等问题,采用结合Bert、GAT、胶囊网络的方法,通过Bert预处理,GAT学习文本上下文特征,然后通过胶囊网络对特征进行主题分类。本发明通过胶囊网络中动态路由的过程,减少了卷积神经网络中数据在反向传播过程中可能给文本分类器模型带来的一些弊端,利用胶囊网络的归一化得出分类的概率,最终也能得到较好的效果。

[0072] 以上为本申请提供的一种文本分类方法的第一个实施例的详细说明,下面为本申请提供的一种文本分类方法的第二个实施例的详细说明。

[0073] 请参阅图2,本申请实施例在上述第一个实施例提供的一种文本分类方法的基础上,在步骤101还包括:

[0074] 步骤201、对文本数据进行预处理,其中预处理包括,去除停用词和去除标点符号。

[0075] 进一步地,还可以包括:

[0076] 步骤200、将预置的训练样本数据输入至初始胶囊网络模型,对初始胶囊网络模型进行训练,得到胶囊网络分类模型,其中,训练样本数据为将预置的样本文本数据通过Bert文本处理模型和上下文特征提取模型处理得到的上下文特征样本数据。

[0077] 以上为本申请提供的一种文本分类方法的第二个实施例的详细说明,下面为本申请提供的一种文本分类装置的第一个实施例的详细说明。

[0078] 请参阅图3,本申请第二方面提供了一种文本分类装置,包括:

[0079] 文本获取单元301,用于获取文本数据;

[0080] 分词处理单元302,用于通过预设的Bert文本处理模型对文本数据进行分词处理,得到词向量;

[0081] 上下文特征提取单元303,用于通过预设的上下文特征提取模型,从文本数据提取各个词向量对应的上下文特征数据;

[0082] 文本分类单元304,用于将上下文特征数据输入至预设的胶囊网络分类模型进行主题分类运算,输出分类结果。

[0083] 进一步地,获取文本数据之后还包括:

[0084] 预处理单元305,用于对文本数据进行预处理,其中预处理包括,去除停用词和去

除标点符号。

[0085] 进一步地,还包括:

[0086] 胶囊网络分类模型训练单元306,用于将预置的训练样本数据输入至初始胶囊网络模型,对初始胶囊网络模型进行训练,得到胶囊网络分类模型,其中,训练样本数据为将预置的样本文本数据通过Bert文本处理模型和上下文特征提取模型处理得到的上下文特征样本数据。

[0087] 以上为本申请提供的一种文本分类装置的第一个实施例的详细说明,下面为本申请提供的一种文本分类方法对应的终端和存储介质的详细说明。

[0088] 本申请第四个实施例提供了一种终端,包括:存储器和处理器;

[0089] 存储器用于存储与本申请第一方面的文本分类方法对应的程序代码;

[0090] 处理器用于执行程序代码。

[0091] 本申请第五个实施例提供了一种存储介质,存储介质中保存有与本申请第一方面的文本分类方法对应的程序代码。

[0092] 所属领域的技术人员可以清楚地了解到,为描述的方便和简洁,上述描述的系统,装置和单元的具体工作过程,可以参考前述方法实施例中的对应过程,在此不再赘述。

[0093] 在本申请所提供的几个实施例中,应该理解到,所揭露的系统,装置和方法,可以通过其它的方式实现。例如,以上所描述的装置实施例仅仅是示意性的,例如,所述单元的划分,仅仅为一种逻辑功能划分,实际实现时可以有另外的划分方式,例如多个单元或组件可以结合或者可以集成到另一个系统,或一些特征可以忽略,或不执行。另一点,所显示或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些接口,装置或单元的间接耦合或通信连接,可以是电性,机械或其它的形式。

[0094] 本申请的说明书及上述附图中的术语“第一”、“第二”、“第三”、“第四”等(如果存在)是用于区别类似的对象,而不必用于描述特定的顺序或先后次序。应该理解这样使用的数据在适当情况下可以互换,以便这里描述的本申请的实施例例如能够以除了在这里图示或描述的那些以外的顺序实施。此外,术语“包括”和“具有”以及他们的任何变形,意图在于覆盖不排他的包含,例如,包含了一系列步骤或单元的过程、方法、系统、产品或设备不必限于清楚地列出的那些步骤或单元,而是可包括没有清楚地列出的或对于这些过程、方法、产品或设备固有的其它步骤或单元。

[0095] 所述作为分离部件说明的单元可以是或者也可以不是物理上分开的,作为单元显示的部件可以是或者也可以不是物理单元,即可以位于一个地方,或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部单元来实现本实施例方案的目的。

[0096] 另外,在本发明各个实施例中的各功能单元可以集成在一个处理单元中,也可以是各个单元单独物理存在,也可以两个或两个以上单元集成在一个单元中。上述集成的单元既可以采用硬件的形式实现,也可以采用软件功能单元的形式实现。

[0097] 所述集成的单元如果以软件功能单元的形式实现并作为独立的产品销售或使用,可以存储在一个计算机可读取存储介质中。基于这样的理解,本发明的技术方案本质上或者说对现有技术做出贡献的部分或者该技术方案的全部或部分可以以软件产品的形式体现出来,该计算机软件产品存储在一个存储介质中,包括若干指令用以使得一台计算机

设备(可以是个人计算机,服务器,或者网络设备等)执行本发明各个实施例所述方法的全部或部分步骤。而前述的存储介质包括:U盘、移动硬盘、只读存储器(ROM,Read-Only Memory)、随机存取存储器(RAM,Random Access Memory)、磁碟或者光盘等各种可以存储程序代码的介质。

[0098] 以上所述,以上实施例仅用以说明本申请的技术方案,而非对其限制;尽管参照前述实施例对本申请进行了详细的说明,本领域的普通技术人员应当理解:其依然可以对前述各实施例所记载的技术方案进行修改,或者对其中部分技术特征进行等同替换;而这些修改或者替换,并不使相应技术方案的本质脱离本申请各实施例技术方案的精神和范围。

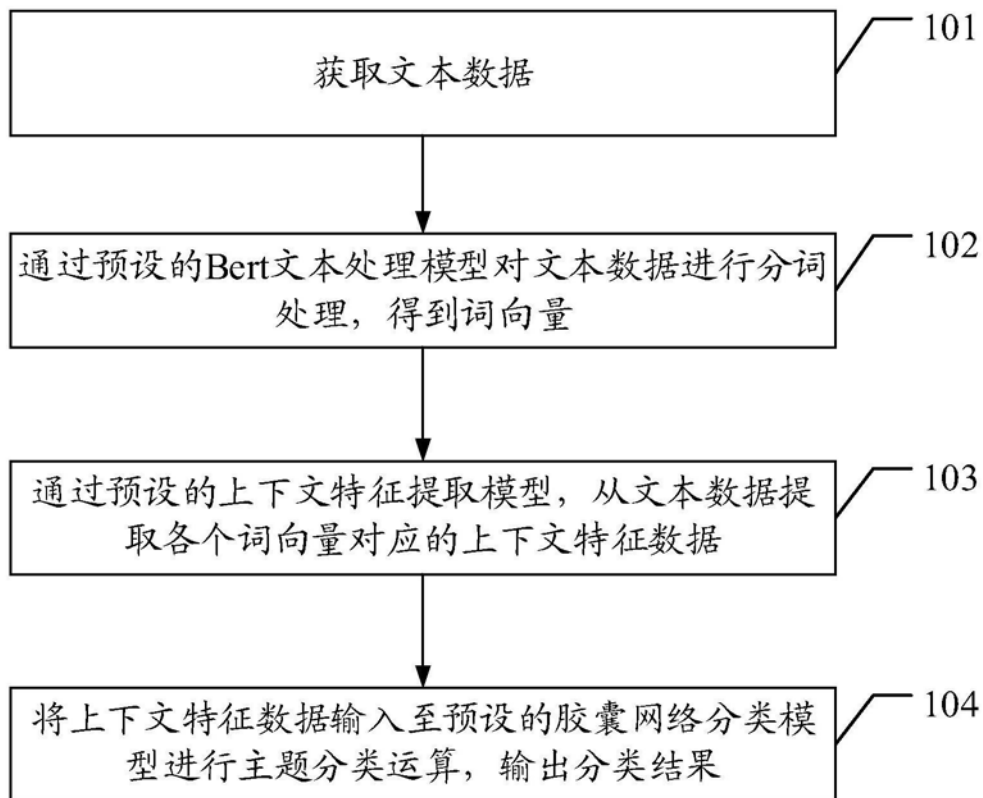


图1

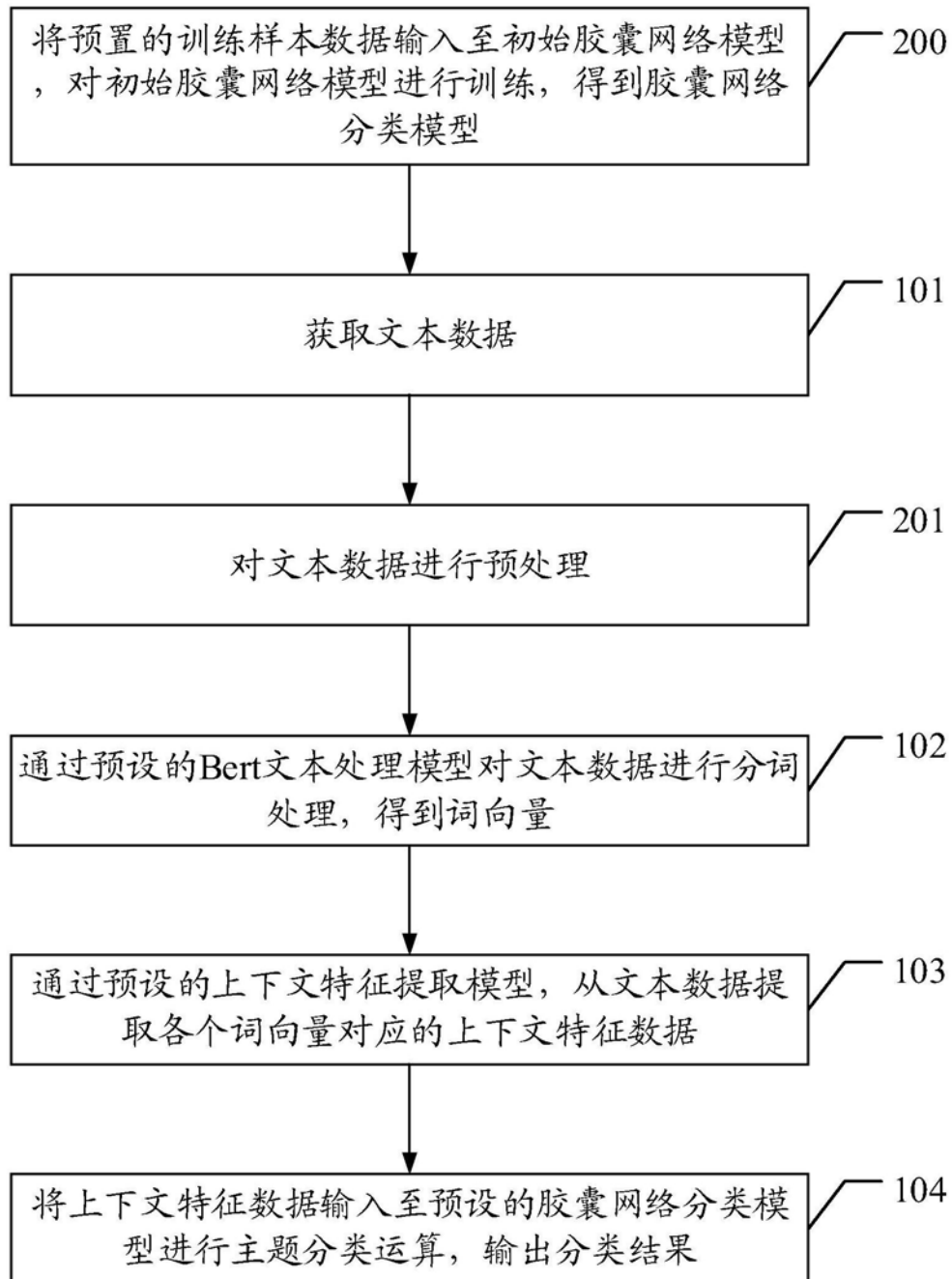


图2

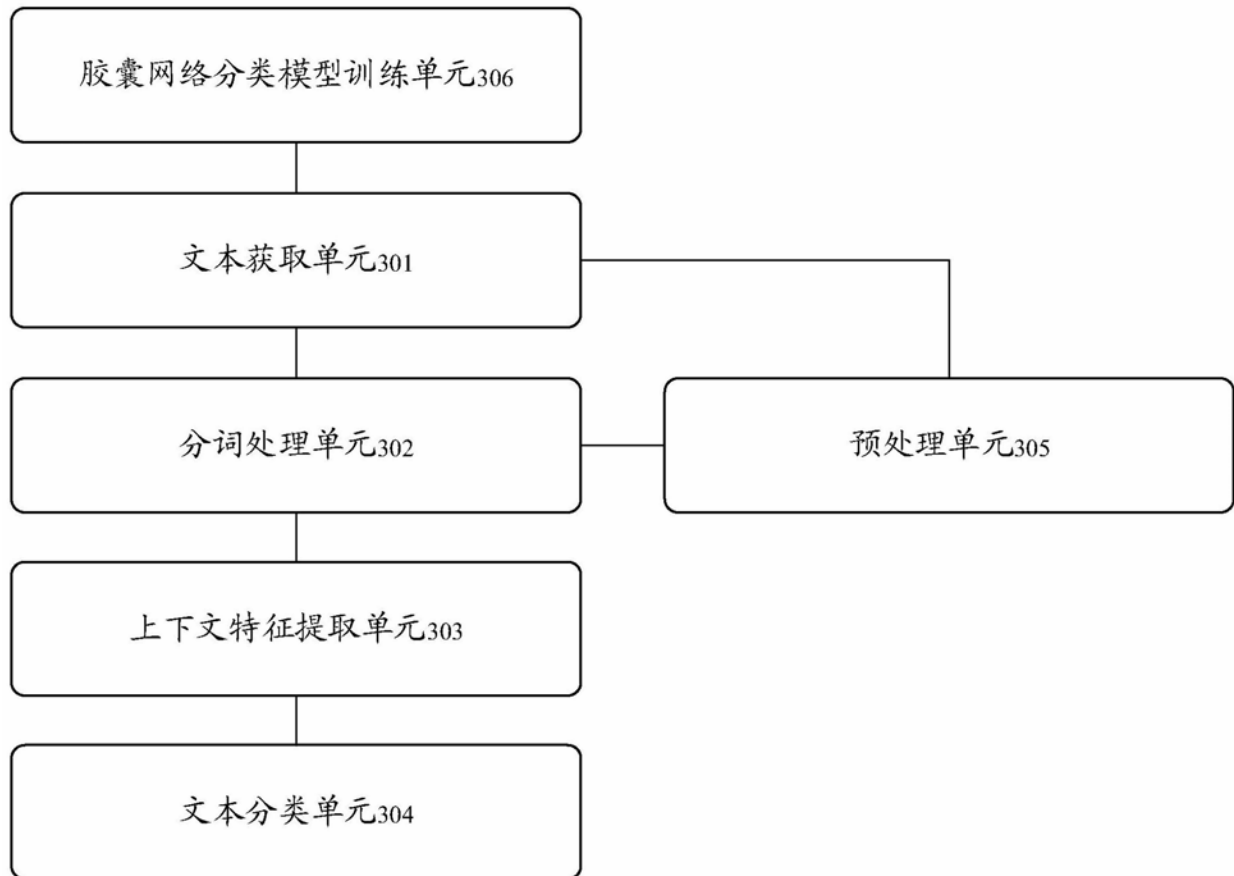


图3

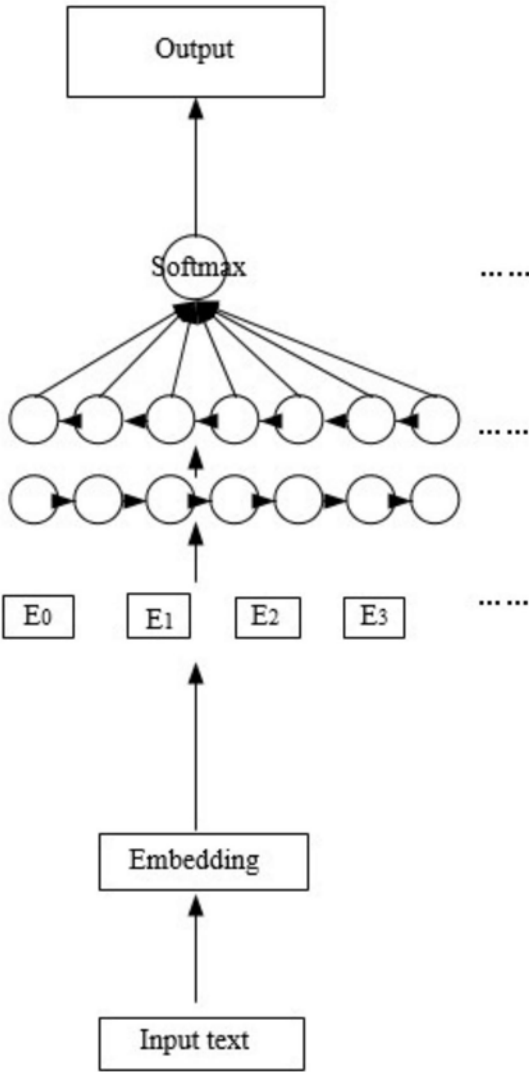


图4

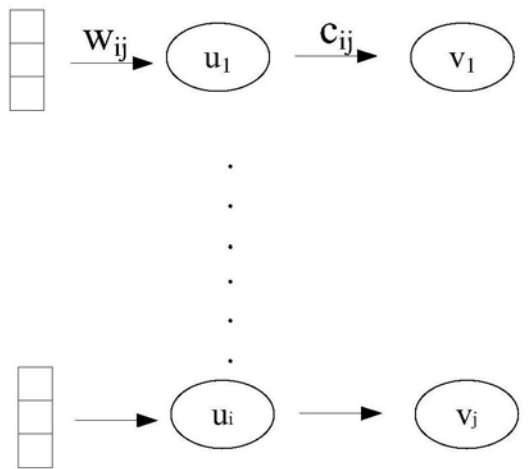


图5