



(12) 发明专利

(10) 授权公告号 CN 111695357 B

(45) 授权公告日 2024. 11. 01

(21) 申请号 202010465811.4

G06F 40/30 (2020.01)

(22) 申请日 2020.05.28

G06N 3/045 (2023.01)

G06N 3/084 (2023.01)

(65) 同一申请的已公布的文献号

申请公布号 CN 111695357 A

(56) 对比文件

(43) 申请公布日 2020.09.22

CN 109034203 A, 2018.12.18

CN 109325112 A, 2019.02.12

(73) 专利权人 平安科技(深圳)有限公司

地址 518000 广东省深圳市福田区福田街
道福安社区益田路5033号平安金融中
心23楼

审查员 李若童

(72) 发明人 李文斌 喻宁 冯晶凌 柳阳

(74) 专利代理机构 广州三环专利商标代理有限
公司 44202

专利代理师 熊永强

(51) Int. Cl.

G06F 16/35 (2019.01)

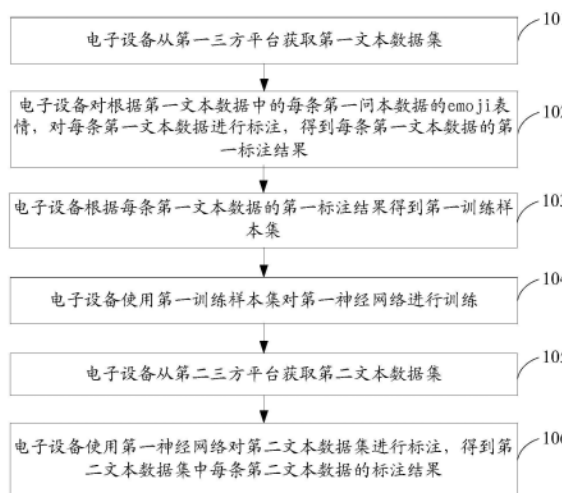
权利要求书2页 说明书11页 附图4页

(54) 发明名称

文本标注方法及相关产品

(57) 摘要

本申请涉及人工智能中的情绪识别技术领域,具体公开了一种文本标注方法及相关产品,该方法包括:从第一三方平台获取第一文本数据集,第一文本数据集中的每条第一文本数据包括emoji表情;根据第一文本数据集中的每条第一文本数据的emoji表情,对每条第一文本数据进行标注,得到每条第一文本数据的第一标注结果,第一标注结果包括正面评价或负面评价;根据每条第一文本数据的第一标注结果得到第一训练样本集;使用第一训练样本集对第一神经网络进行训练;从第二三方平台获取第二文本数据集;使用第一神经网络对第二文本数据集进行标注,得到第二文本数据集中每条第二文本数据的第二标注结果,第二标注结果包括正面评价、负面评价或中性评价中的一种。



1. 一种文本标注方法,其特征在于,应用于电子设备,包括:

所述电子设备从第一三方平台获取第一文本数据集,所述第一文本数据集中的每条第一文本数据包括emoji表情;

所述电子设备对所述第一文本数据集中的每条第一文本数据进行清洗,删除不包含emoji表情的第一文本数据,得到新的第一文本数据集;将所述新的第一文本数据集作为所述第一文本数据集;

所述电子设备根据所述第一文本数据集中的每条第一文本数据的emoji表情,对每条第一文本数据进行标注,得到每条第一文本数据的第一标注结果,所述第一标注结果包括正面评价或负面评价;包括:

根据所述第一文本数据集中的每条第一文本数据的emoji表情,确定每条第一文本数据的第一情感评价,所述第一情感评价包括正面评价或负面评价;提取每条第一文本数据的文本内容;对每条第一文本数据的文本内容进行语义分析,得到每条第一文本数据的语义信息;根据每条第一文本数据的语义信息,确定每条第一文本数据的第二情感评价;保留所述第一文本数据集中第一情感评价和第二情感评价一致的第一文本数据,删除第一情感评价和第二情感评价不一致的第一文本数据;根据每条第一文本数据的第一情感评价,对每条第一文本数据进行标注;

所述电子设备根据每条第一文本数据的第一标注结果得到第一训练样本集;

所述电子设备使用所述第一训练样本集对第一神经网络进行训练;

所述电子设备从第二三方平台获取第二文本数据集;

所述电子设备使用所述第一神经网络对所述第二文本数据集进行标注,得到所述第二文本数据集中每条第二文本数据的第二标注结果,所述第二标注结果包括正面评价、负面评价或中性评价中的一种;

所述电子设备根据所述第二文本数据集中每条第二文本数据的第二标注结果,得到第二训练样本集;所述电子设备使用所述第二训练样本集对第二神经网络进行训练;所述电子设备获取任意一条待发表的评论数据;所述电子设备使用所述第二神经网络对所述待发表的评论数据进行情感分类,得到对所述待发表的评论数据的分类结果;所述电子设备根据所述分类结果,确定是否公开所述待发表的评论数据。

2. 根据权利要求1所述的方法,其特征在于,所述使用所述第一神经网络对所述第二文本数据集进行标注,得到所述第二文本数据集中每条第二文本数据的第二标注结果,包括:

使用所述第一神经网络对所述第二文本数据集中的每条第二文本数据进行分类,得到每条第二文本数据为正面评价的第一概率和负面评价的第二概率;

确定第一概率大于第一阈值的第二文本数据的第二标注结果为正面评价;

确定第二概率大于所述第一阈值的第二文本数据的第二标注结果为负面评价;

确定第一概率小于所述第一阈值,且大于第二阈值的第二文本数据的第二标注结果为中性评价。

3. 根据权利要求1所述的方法,其特征在于,所述电子设备根据所述第二文本数据集中每条第二文本数据的第二标注结果,得到第二训练样本集之后,所述方法还包括:

将所述第二训练样本与所述第一训练样本集进行合并,得到新的第二训练样本集;

所述电子设备使用所述第二训练样本集对第二神经网络进行训练,包括:

所述电子设备使用所述新的第二训练样本集对第二神经网络进行训练。

4. 一种电子设备,其特征在於,所述电子设备用于执行权利要求1-3任一项所述的方法,所述电子设备包括:

获取单元,用于从第一三方平台获取第一文本数据集,所述第一文本数据集中的每条第一文本数据包括emoji表情;

标注单元,根据所述第一文本数据集中的每条第一文本数据的emoji表情,对每条第一文本数据进行标注,得到每条第一文本数据的第一标注结果,所述第一标注结果包括正面评价或负面评价;

训练单元,用于根据每条第一文本数据的第一标注结果得到第一训练样本集,并使用所述第一训练样本集对第一神经网络进行训练;

所述获取单元,还用于从第二三方平台获取第二文本数据集;

所述标注单元,还用于使用所述第一神经网络对所述第二文本数据集进行标注,得到所述第二文本数据集中每条第二文本数据的第二标注结果,所述第二标注结果包括正面评价、负面评价或中性评价中的一种。

5. 一种电子设备,其特征在於,包括处理器、存储器、通信接口以及一个或多个程序,其中,所述一个或多个程序被存储在所述存储器中,并且被配置由所述处理器执行,所述程序包括用于执行如权利要求1-3任一项所述的方法中的步骤的指令。

6. 一种计算机可读存储介质,其特征在於,所述计算机可读存储介质存储有计算机程序,所述计算机程序被处理器执行以实现如权利要求1-3任一项所述的方法。

文本标注方法及相关产品

技术领域

[0001] 本申请涉及人工智能中的情绪识别技术领域,具体涉及一种文本标注方法及相关产品。

背景技术

[0002] 随着人工智能的发展,神经网络应用的范围越来越广泛。例如,在视频监控领域,可使用神经网络对监控视频中的人物识别或者在医疗领域,使用神经网络对核磁共振图像中肿瘤进行识别;再者,在文字识别领域,使用神经网络对文本进行感情分类。

[0003] 虽然神经网络对图像识别有着不错的表现。但是,前期对神经网络的训练需要数量足够多,质量足够高的训练数据集。而训练数据集的制作是一个成本非常高的项目。首先需要从数据库中获取一些质量较高的原始数据集,并对该原始数据集进行标注。例如,训练文本情感分类网络时,需要获取大量语义完整,情感明确的文本,然后,人工对该大量的文本进行标注。然而,由于文本的数量极其庞大,人工标注需要投入大量时间和人力成本,标注效率低。

发明内容

[0004] 本申请实施例提供了一种文本标注方法及相关产品。增加了对文本标注的应用场景,以及提高对文本标注的效率。

[0005] 第一方面,本申请实施例提供一种文本标注方法,应用于电子设备,包括:

[0006] 所述电子设备从第一三方平台获取第一文本数据集,所述第一文本数据集中的每条第一文本数据包括emoji表情;

[0007] 所述电子设备根据所述第一文本数据集中的每条第一文本数据的emoji表情,对每条第一文本数据进行标注,得到每条第一文本数据的第一标注结果,所述第一标注结果包括正面评价或负面评价;

[0008] 所述电子设备根据每条第一文本数据的第一标注结果得到第一训练样本集;

[0009] 所述电子设备使用所述第一训练样本集对第一神经网络进行训练;

[0010] 所述电子设备从第二三方平台获取第二文本数据集;

[0011] 所述电子设备使用所述第一神经网络对所述第二文本数据集进行标注,得到所述第二文本数据集中每条第二文本数据的第二标注结果,所述第二标注结果包括正面评价、负面评价或中性评价中的一种。

[0012] 第二方面,本申请实施例提供一种电子设备,包括:

[0013] 获取单元,用于从第一三方平台获取第一文本数据集,所述第一文本数据集中的每条第一文本数据包括emoji表情;

[0014] 标注单元,根据所述第一文本数据集中的每条第一文本数据的emoji表情,对每条第一文本数据进行标注,得到每条第一文本数据的第一标注结果,所述第一标注结果包括正面评价或负面评价;

[0015] 训练单元,用于根据每条第一文本数据的第一标注结果得到第一训练样本集,并使用所述第一训练样本集对第一神经网络进行训练;

[0016] 所述获取单元,还用于从第二三方平台获取第二文本数据集;

[0017] 所述标注单元,还用于使用所述第一神经网络对所述第二文本数据集进行标注,得到所述第二文本数据集中每条第二文本数据的第二标注结果,所述第二标注结果包括正面评价、负面评价或中性评价中的一种。

[0018] 第三方面,本申请实施例提供一种电子设备,包括处理器、存储器、通信接口以及一个或多个程序,其中,所述一个或多个程序被存储在所述存储器中,并且被配置由所述处理器执行,所述程序包括用于执行如第一方面所述的方法中的步骤的指令。

[0019] 第四方面,本申请实施例提供一种计算机可读存储介质,所述计算机可读存储介质存储有计算机程序,所述计算机程序使得计算机执行如第一方面所述的方法。

[0020] 第五方面,本申请实施例提供一种计算机程序产品,所述计算机程序产品包括存储了计算机程序的非瞬时性计算机可读存储介质,所述计算机可操作来使计算机执行如第一方面所述的方法。

[0021] 实施本申请实施例,具有如下有益效果:

[0022] 可以看出,在本申请实施例中,通过文本数据中的emoji表情对评论数据进行标注,无需对评论数据进行语义分析,从而在标注时不会受文本数据的语言类型的限制,增加了该文本标注的应用场景;另外,可通过emoji表情对文本数据进行自动标注,无需人工标注,节省了人力物力资源。

附图说明

[0023] 为了更清楚地说明本申请实施例中的技术方案,下面将对实施例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的前提下,还可以根据这些附图获得其他的附图。

[0024] 图1为本申请实施例提供的一种标注方法的流程示意图;

[0025] 图2为本申请实施例提供的另一种标注方法的流程示意图;

[0026] 图3为本申请实施例提供的另一种标注方法的流程示意图;

[0027] 图4为本申请实施例提供的一种电子设备的结构示意图;

[0028] 图5为本申请实施例提供的一种电子设备的功能单元组成框图。

具体实施方式

[0029] 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作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。

[0030] 本申请的说明书和权利要求书及所述附图中的术语“第一”、“第二”、“第三”和“第四”等是用于区别不同对象,而不是用于描述特定顺序。此外,术语“包括”和“具有”以及它们任何变形,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、产品或设备没有限定于已列出的步骤或单元,而是可选地还包括没有列出的步骤或单元,或可选地还包括对于这些过程、方法、产品或设备固有的其它步骤或单元。

[0031] 在本文中提及“实施例”意味着,结合实施例描述的特定特征、结果或特性可以包含在本申请的至少一个实施例中。在说明书中的各个位置出现该短语并不一定均是指相同的实施例,也不是与其它实施例互斥的独立的或备选的实施例。本领域技术人员显式地和隐式地理解的是,本文所描述的实施例可以与其它实施例相结合。

[0032] 本申请中的电子设备可以包括智能手机(如Android手机、iOS手机、Windows Phone手机等)、平板电脑、掌上电脑、笔记本电脑、移动互联网设备MID(Mobile Internet Devices,简称:MID)或穿戴式设备等。上述电子设备仅是举例,而非穷举,包括但不限于上述电子设备。在实际应用中,上述电子设备还可以包括:智能车载终端、计算机设备等等。

[0033] 参阅图1,图1为本申请实施例提供的一种文本标注方法的流程示意图,该方法应用于电子设备,该方法包括以下步骤内容:

[0034] 101:电子设备从第一三方平台获取第一文本数据集。

[0035] 其中,该第一三方平台可以为微博、twitter、Facebook,等社交应用或Amazon淘宝京东,等电商平台。即该第一三方平台为包含正面评价的文本数据和负面评价的文本数据较多的第三方平台。电子设备通过该第一三方平台提供的应用程序接口(Application Programming Interface,API)从该第一平台中随机多条第一文本数据,得到第一文本数据集。即电子设备遵从第一三方平台的Robot协议,通过该第一三方平台的API从第一三方平台中获得第一文本数据集。



[0036] 在一些可能的实施方式中,由于第一文本数据是通过第一三方平台的API中获取的,未进行人工审核,有些第一文本数据可能并不符合要求。例如,不包含emoji表情或者文本内容过短。因此,在得到多条第一文本数据后,先对该第一文本数据集中的第一文本数据进行清洗,以清洗掉不包含emoji表情或者文本内容过短的第一文本数据,将清洗后的第一文本数据组成该第一文本数据集。

[0037] 因此,该第一文本数据集中的每条第一文本数据包含有emoji表情。

[0038] 102:电子设备对根据第一文本数据中的每条第一问本数据的emoji表情,对每条第一文本数据进行标注,得到每条第一文本数据的第一标注结果。

[0039] 其中,该第一标注结果包括证明评价或负面评价。

[0040] 示例性的,对该第一文本数据进行清洗,则该第一文本数据集中的每条第一文本


数据包括emoji表情。由于,emoji表情本身携带情感评价。例如,emoji表情 、

表示的情感评价为正面评价,而Emoji表情 、表示负面评价。因此可根据每

条第一文本数据的emoji表情确定每条第一文本数据的第一情感评价;然后,根据每条第一文本数据的第一情感评价对每条第一文本数据进行标注,即为每条第一文本数据添加情感标签。即在任意一条第一文本数据的emoji表情属于正面评价的emoji表情集合的情况下,则将该第一文本数据标注为正面评价,在该emoji表情属于负面评价的emoji表情集合的情况下,则将该第一文本数据标注为负面评价。

[0041] 其中,与该第一文本数据相对应的,该第一标注结果包括正面评价和负面评价,该正面评价对应的情感包括开心、赞同、欣赏,等情感,负面评价对应的情感包括愤怒、悲观、不赞成,等情感。

[0042] 需要说明,有些emoji表情并没有把握确定出该emoji表情对应的情感评价。例如,

emoji表情  可以表示开心,即正面感情,也可以用来表示嘲讽,即负面情感。对于不

对该第一文本数据集中包含有这些emoji表情的第一文本数据进行标注,只对包含有正面评价对应的emoji表情或者负面评价对应的emoji表情的第一文本数据进行标注。

[0043] 进一步地,为了提高通过emoji表情标注的精确度,可提取每条第一文本数据的文本内容,对每条第一文本数据的文本内容进行语义分析,得到每条第一文本数据的语义信息;根据每条第一文本数据的语义信息,确定每条第一文本数据的第一情感评价;保留第一文本数据集中第一情感评价和第二情感评价一致的第一文本数据,删除第一情感评价和第二情感评价不一致的第一文本数据。通过语义分析与emoji表情进行双重标注,从而降低单方面通过emoji表情标注带来的标注误差,提高对第一文本数据集标注的精确度。

[0044] 103:电子设备根据每条第一文本数据的第一标注结果得到第一训练样本集。

[0045] 即将标注好的第一文本数据作为带有标签的训练样本,得到该第一训练样本集合。

[0046] 104:电子设备使用第一训练样本集对第一神经网络进行训练。

[0047] 具体来说,先构建第一神经网络的初始参数,将该第一训练样本集中的训练样本输入到该第一神经网络,得到对该训练样本的预测结果;然后,基于该预测结果和该训练样本的标注结果确定损失梯度,基于该损失梯度构造损失函数;最后,基于该损失函数以及梯度下降法反向更新该初始参数的参数值;直到该第一神经网络收敛,完成对该第一神经网络的训练。

[0048] 105:电子设备从第二三方平台获取第二文本数据集。

[0049] 其中,该第二三方平台可以为发表科技类新闻或wiki或summary文本的新闻平台。即该第二三方平台为包含有大量的中性评价的文本数据的三方平台。

[0050] 同样,电子设备遵从第二三方平台的Robot协议,通过该第二三方平台的API从第二三方平台中获取多条第二文本数据,得到该第二文本数据集。

[0051] 当然,在获取多条第二文本数据后,可对该多条第二文本数据进行清洗,以清洗掉不合法、文本内容过短的第二文本数据。

[0052] 106:电子设备使用第一神经网络对第二文本数据集进行标注,得到第二文本数据集中每条第二文本数据的标注结果。

[0053] 其中,该第二标注结果包括正面评价、负面评价或中性评价中的一种。

[0054] 具体地,电子设备使用第一神经网络,对第二文本数据集中的每条第二文本数据进行分类,得到每条第二文本数据为正面评价的第一概率和负面评价的第二概率;然后,将第一概率大于第一阈值(即有100%把握认为该第二文本数据的情感评价为正面评价)的第二文本数据标注为正面评价;将第二概率大于该第一阈值(即有100%把握认为该第二文本数据的情感评价为负面评价)的第二文本数据标注为负面评价;将第一概评价的训练样本

率小于所述第一阈值,且大于第二阈值(即没有100%把握认为该第二文本数据的情感评价为正面评价还是负面评价)的第二文本数据标注为中性评价。

[0055] 其中,该第一阈值可以为0.7、0.75、0.8或者其他值。该第二阈值可以为0.4、0.45、0.5或者其他值。

[0056] 可以看出,在本申请实施例中,通过文本数据中的emoji表情对文本数据进行标注,无需对文本数据进行语义分析,从而在标注时不会受文本数据的语言类型的限制,进而增加了该标注方法的应用场景;另外,可通过emoji表情对文本数据进行自动标注,无需人工标注即可完成对文本数据进行标注,从而节省了人力物力资源。

[0057] 在一些可能的实施方式中,所述方法还包括:

[0058] 电子设备根据该第二文本数据集中每条第二文本数据的第二标注结果,得到第二训练样本集,即根据第二文本数据集中每条第二文本数据的标注结果,将该第二文本数据集组成有标签的第二训练样本集;然后,使用该第二训练样本集对第二神经网络进行训练;并获取任意一条待发表的评论数据,使用第二神经网络对所述待发表的评论数据进行分类,得到对所述待发表的评论数据的分类结果;根据所述分类结果确定是否公开所述待发表的评论数据。

[0059] 其中,在该待发表的评论数据可以为任意一个新闻网站下的待发表的评论数据的情况下,则当该分类结果为正面评价或者中性评价时,则公开该待发表的评论数据,当该分类结果为负面评价时,则不公开该待发表的评论数据。相比现有的通过人工审核待发表的评论数据,本申请中可以通过第二神经网络自动对该待发表的评论数据进行审核,进而节省了人力资源。

[0060] 其中,在该待发表的评论数据可以为任意一个电商平台下的评论数据的情况下,则当该分类结果为正面评价或者负面评价时,则将该待发表的评论数据与用户的购买记录进行核对,确定该待发表的评论数据的真实性,在确定该待发表的评论数据为恶意刷屏的情况下,则不公开该待发表的评论数据。本申请中可以通过第二神经网络自动对该待发表的评论数据进行审核,以确定该待发表的评论数据的真实性,进而节省了人力资源。

[0061] 在一些可能的实施方式中,由于从该第二三方平台中获得的第二文本数据大多数为中性文本数据,而从第一三方平台中获得的第一文本数据大多数为正面评价的文本数据和负面评价的文本数据。因此,为了增加第二训练样本集中正面评价的训练样本和负面评价的训练样本的数量,可将第二训练样本集与该第一训练样本集进行合并,得到训练样本充足的新的第二训练样本集,使用该新的第二训练样本集进行对第二神经网络进行训练,进而使训练出的第二神经网络更加精确。

[0062] 在一些可能的实施方式中,在根据所述第一文本数据集中的每条第一文本数据的emoji表情,确定每条第一文本数据的第一情感评价之后,所述方法还包括:

[0063] 提取每条第一文本数据的文本内容;将所述文本内容转化为第二emoji表情;根据所述第二emoji表情确定每条第一文本数据对应的第二情感评价;确定每条第一文本数据的第一情感评价与第二情感评价是否一致,若一致,则根据每条第一文本数据的第一情感评价,对每条第一文本数据进行标注。通过文本转emoji操作,对每条第一文本数据对应的情感评价进行验证,进而提高后续对第一文本数据标注的精确度。

[0064] 在一些可能的实施方式中,所述方法还包括:

[0065] 获取任意用户的评论数据,该评论数据为所述用户对目标产品的评论数据,该目标产品包括理财产品;使用上述第二神经网络对所述用户的评论数据进行分类,得到对所述用户的评论数据的分类结果;根据所述用户的评论数据的分类结果,筛选目标用户,即将分类结果为正面评价的用户作为目标用户;向所述目标用户推荐所述目标产品。

[0066] 可以看出,在本实施例中,使用第二神经网络筛选出对目标产品(理财产品)感兴趣的,保证用户筛选的精确性,提高推荐的成功率。

[0067] 参阅图2,图2为本申请实施例提供的另一种文本标注方法的流程示意图该实施例中与图1所示的实施例相同的内容,此处不再重复描述。该方法应用于电子设备,该方法包括以下步骤内容:

[0068] 201:电子设备从第一平台获取第一文本数据集。

[0069] 202:电子设备对所述第一文本数据集中的每条第一文本数据进行清洗,删除不包含emoji表情的第一文本数据,得到新的第一文本数据集,将所述新的第一文本数据集作为所述第一文本数据集。

[0070] 203:电子设备根据所述第一文本数据集中的每条第一文本数据的emoji表情,确定每条第一文本数据的第一情感评价,该第一情感评价包括正面评价或负面评价。

[0071] 204:电子设备提取每条第一文本数据的文本内容,对每条第一文本数据的文本内容进行语义分析,得到每条第一文本数据的语义信息。

[0072] 205:电子设备根据每条第一文本数据的语义信息,确定每条第一文本数据的第二情感评价。

[0073] 206:电子设备保留所述第一文本数据集中第一情感评价和第二情感评价一致的第一文本数据,删除第一情感评价和第二情感评价不一致的第一文本数据。

[0074] 207:电子设备根据剩余的第一文本数据的第一情感评价,对该剩余的第一文本数据进行标注,得到第一训练样本集。

[0075] 其中,该剩余的第一文本数据为该第一文本数据集中删除第一情感评价和第二情感评价不一致的第一评论数据之后剩余的第一文本数据。

[0076] 208:电子设备使用第一训练样本集对第一神经网络进行训练。

[0077] 209:电子设备从第二平台获取第二文本数据集。

[0078] 210:电子设备使用所述第一神经网络对所述第二文本数据集进行标注,得到所述第二文本数据集中每条第二文本数据的第二标注结果,所述第二标注结果包括正面评价、负面评价或中性评价中的一种。

[0079] 可以看出,在本申请实施例中,通过评论数据中的emoji表情对评论数据进行标注,无需对评论数据进行语义分析,从而在标注时不会受评论数据的语言类型的限制,进而增加了该标注方法的应用场景;另外,可通过emoji表情对评论数据进行自动标注,无需人工标注即可得到包含有情感分类标签的训练样本集,从而节省了人力物力资源;而且,在对第一文本数据集进行标注前,先对第一文本数据集进行清洗保留高质量的第一文本数据,从而提高了标注的精确度。

[0080] 参阅图3,图3为本申请实施例提供的另一种文本标注方法的流程示意图该实施例中与图1和图2所示的实施例相同的内容,此处不再重复描述。该方法应用于电子设备,该方法包括以下步骤内容:

- [0081] 301:电子设备从第一平台获取第一文本数据集。
- [0082] 302:电子设备对所述第一文本数据集中的每条第一文本数据进行清洗,删除不包含emoji表情的第一文本数据,得到新的第一文本数据集,将所述新的第一文本数据集作为所述第一文本数据集。
- [0083] 303:电子设备根据所述第一文本数据集中的每条第一文本数据的emoji表情,确定每条第一文本数据的第一情感评价,该第一情感评价包括正面评价或负面评价。
- [0084] 304:电子设备提取每条第一文本数据的文本内容,对每条第一文本数据的文本内容进行语义分析,得到每条第一文本数据的语义信息。
- [0085] 305:电子设备根据每条第一文本数据的语义信息,确定每条第一文本数据的第二情感评价。
- [0086] 306:电子设备保留所述第一文本数据集中第一情感评价和第二情感评价一致的第一文本数据,删除第一情感评价和第二情感评价不一致的第一文本数据。
- [0087] 307:电子设备根据剩余的第一文本数据的第一情感评价,对该剩余的第一文本数据进行标注,得到第一训练样本集。
- [0088] 其中,该剩余的第一文本数据为该第一文本数据集中删除第一情感评价和第二情感评价不一致的第一评论数据之后剩余的第一文本数据。
- [0089] 308:电子设备使用第一训练样本集对第一神经网络进行训练。
- [0090] 309:电子设备从第二平台获取第二文本数据集。
- [0091] 310:电子设备使用所述第一神经网络对所述第二文本数据集进行标注,得到所述第二文本数据集中每条第二文本数据的第二标注结果,所述第二标注结果包括正面评价、负面评价或中性评价中的一种。
- [0092] 311:电子设备使用根据每条第二文本数据的第二标注结果,得到第二训练样本集,并使用该第二训练样本集对第二神经网络进行训练。
- [0093] 312:电子设备获取任意一条评论数据,使用所述第二神经网络对所述评论数据进行分类,得到对所述评论数据的分类结果,根据所述分类结果,确定是否公开所述评论数据。
- [0094] 可以看出,在本申请实施例中,通过文本数据中的emoji表情对文本数据进行标注,无需对文本数据进行语义分析,从而在标注时不会受文本数据的语言类型的限制,进而增加了该标注方法的应用场景;另外,可通过emoji表情对文本数据进行自动标注,无需人工标注即可得到包含有情感分类标签的训练样本集,从而节省了人力物力资源;而且,在对第一文本数据集进行标注前,先对第一文本数据集进行清洗保留高质量的第一文本数据,从而提高了标注的精确度;此外,使用训练好的第二神经网络对待发表的评论数据进行分类,自动屏蔽不满足要求的待发表的评论数据,无需人力审核,节省了人力资源。
- [0095] 参阅图4,图4为本申请实施例提供的一种电子设备的结构示意图。如图4所示,电子设备400包括处理器、存储器、通信接口以及一个或多个程序,其中,上述一个或多个程序被存储在上述存储器中,并且被配置由上述处理器执行,上述程序包括用于执行以下步骤的指令:
- [0096] 从第一三方平台获取第一文本数据集,所述第一文本数据集中的每条第一文本数据包括emoji表情;

[0097] 根据所述第一文本数据集中的每条第一文本数据的emoji表情,对每条第一文本数据进行标注,得到每条第一文本数据的第一标注结果,所述第一标注结果包括正面评价或负面评价;

[0098] 根据每条第一文本数据的第一标注结果得到第一训练样本集;

[0099] 使用所述第一训练样本集对第一神经网络进行训练;

[0100] 从第二三方平台获取第二文本数据集;

[0101] 使用所述第一神经网络对所述第二文本数据集进行标注,得到所述第二文本数据集中每条第二文本数据的第二标注结果,所述第二标注结果包括正面评价、负面评价或中性评价中的一种。

[0102] 在一些可能的实施方式中,在根据所述第一文本数据集中的每条第一文本数据的emoji表情,对每条第一文本数据进行标注方面,上述程序具体用于执行以下步骤的指令:

[0103] 根据所述第一文本数据集中的每条第一文本数据的emoji表情,确定每条第一文本数据的第一情感评价,所述第一情感评价包括正面评价或负面评价;

[0104] 根据每条第一文本数据的第一情感评价,对每条第一文本数据进行标注。

[0105] 在一些可能的实施方式中,根据所述第一文本数据集中的每条第一文本数据的emoji表情,确定每条第一文本数据的第一情感评价之后,上述程序还用于执行以下步骤的指令:

[0106] 提取每条第一文本数据的文本内容;

[0107] 对每条第一文本数据的文本内容进行语义分析,得到每条第一文本数据的语义信息;

[0108] 根据每条第一文本数据的语义信息,确定每条第一文本数据的第二情感评价;

[0109] 保留所述第一文本数据集中第一情感评价和第二情感评价一致的第一文本数据,删除第一情感评价和第二情感评价不一致的第一文本数据。

[0110] 在一些可能的实施方式中,在对所述第一文本数据集进行标注之前,上述程序还用于执行以下步骤的指令:

[0111] 对所述第一文本数据集中的每条第一文本数据进行清洗,删除不包含emoji表情的第一文本数据,得到新的第一文本数据集;

[0112] 将所述新的第一文本数据集作为所述第一文本数据集。

[0113] 在一些可能的实施方式中,在用所述第一神经网络对所述第二文本数据集进行标注,得到所述第二文本数据集中每条第二文本数据的第二标注结果方面,上述程序具体用于执行以下步骤的指令:

[0114] 使用所述第一神经网络对所述第二文本数据集中的每条第二文本数据进行分类,得到每条第二文本数据为正面评价的第一概率和负面评价的第二概率;

[0115] 确定第一概率大于第一阈值的第二文本数据的第二标注结果为正面评价;

[0116] 确定第二概率大于所述第一阈值的第二文本数据的第二标注结果为负面评价;

[0117] 将第一概率小于所述第一阈值,且大于所述第二阈值的第二文本数据的第二标注结果为中性评价。

[0118] 在一些可能的实施方式中,上述程序还用于执行以下步骤的指令:

[0119] 根据所述第二文本数据集中每条第二文本数据的第二标注结果,得到第二训练样

本集；

[0120] 使用所述第二训练样本集对第二神经网络进行训练；

[0121] 所述电子设备获取任意一条待发表的评论数据；

[0122] 使用所述第二神经网络对所述待发表的评论数据进行情感分类,得到对所述待发表的评论数据的分类结果；

[0123] 根据所述分类结果,确定是否公开所述待发表的评论数据。

[0124] 在一些可能的实施方式中,在根据所述第二文本数据集中每条第二文本数据的第二标注结果,得到第二训练样本集之后,上述程序还用于执行以下步骤的指令：

[0125] 将所述第二训练样本与所述第一训练样本集进行合并,得到新的第二训练样本集；

[0126] 在使用所述第二训练样本集对第二神经网络进行训练方面,上述程序具体用于执行以下步骤的指令：

[0127] 使用所述新的第二训练样本集对第二神经网络进行训练。

[0128] 参阅图5,图5本申请实施例提供一种电子设备的功能单元组成框图。电子设备500包括:获取单元510、标注单元520和训练单元530;其中：

[0129] 获取单元510,用于从第一三方平台获取第一文本数据集,所述第一文本数据集中的每条第一文本数据包括emoji表情；

[0130] 标注单元520,根据所述第一文本数据集中的每条第一文本数据的emoji表情,对每条第一文本数据进行标注,得到每条第一文本数据的第一标注结果,所述第一标注结果包括正面评价或负面评价；

[0131] 训练单元530,用于根据每条第一文本数据的第一标注结果得到第一训练样本集,并使用所述第一训练样本集对第一神经网络进行训练；

[0132] 获取单元510,还用于从第二三方平台获取第二文本数据集；

[0133] 标注单元520,还用于使用所述第一神经网络对所述第二文本数据集进行标注,得到所述第二文本数据集中每条第二文本数据的第二标注结果,所述第二标注结果包括正面评价、负面评价或中性评价中的一种。

[0134] 在一些可能的实施方式中,在根据所述第一文本数据集中的每条第一文本数据的emoji表情,对每条第一文本数据进行标注方面,标注单元520,具体用于：

[0135] 根据所述第一文本数据集中的每条第一文本数据的emoji表情,确定每条第一文本数据的第一情感评价,所述第一情感评价包括正面评价或负面评价；

[0136] 根据每条第一文本数据的第一情感评价,对每条第一文本数据进行标注。

[0137] 在一些可能的实施方式中,电子设备500还包括清洗单元540,根据所述第一文本数据集中的每条第一文本数据的emoji表情,确定每条第一文本数据的第一情感评价之后,清洗单元540,用于：

[0138] 提取每条第一文本数据的文本内容；

[0139] 对每条第一文本数据的文本内容进行语义分析,得到每条第一文本数据的语义信息；

[0140] 根据每条第一文本数据的语义信息,确定每条第一文本数据的第二情感评价；

[0141] 保留所述第一文本数据集中第一情感评价和第二情感评价一致的第一文本数据,

删除第一情感评价和第二情感评价不一致的第一文本数据。

[0142] 在一些可能的实施方式中,电子设备500还包括清洗单元540,在对所述第一文本数据集进行标注之前,清洗单元540用于:

[0143] 对所述第一文本数据集中的每条第一文本数据进行清洗,删除不包含emoji表情的第一文本数据,得到新的第一文本数据集;

[0144] 将所述新的第一文本数据集作为所述第一文本数据集。

[0145] 在一些可能的实施方式中,在使用所述第一神经网络对所述第二文本数据集进行标注,得到所述第二文本数据集中每条第二文本数据的第二标注结果方面,标注单元520,具体用于:

[0146] 使用所述第一神经网络对所述第二文本数据集中的每条第二文本数据进行分类,得到每条第二文本数据为正面评价的第一概率和负面评价的第二概率;

[0147] 确定第一概率大于第一阈值的第二文本数据的第二标注结果为正面评价;

[0148] 确定第二概率大于所述第一阈值的第二文本数据的第二标注结果为负面评价;

[0149] 将第一概率小于所述第一阈值,且大于所述第二阈值的第二文本数据的第二标注结果为中性评价。

[0150] 在一些可能的实施方式中,还包括确定单元550;

[0151] 训练单元530,还用于根据所述第二文本数据集中每条第二文本数据的第二标注结果,得到第二训练样本集;

[0152] 训练单元530,还用于使用所述第二训练样本集对第二神经网络进行训练;

[0153] 确定单元550,用于获取任意一条待发表的评论数据;使用所述第二神经网络对所述待发表的评论数据进行情感分类,得到对所述待发表的评论数据的分类结果;根据所述分类结果,确定是否公开所述待发表的评论数据。

[0154] 在一些可能的实施方式中,在根据所述第二文本数据集中每条第二文本数据的第二标注结果,得到第二训练样本集之后,训练单元530,还用于:

[0155] 将所述第二训练样本与所述第一训练样本集进行合并,得到新的第二训练样本集;

[0156] 在使用所述第二训练样本集对第二神经网络进行训练方面,训练单元530,具体用于:

[0157] 使用所述新的第二训练样本集对第二神经网络进行训练。

[0158] 本申请实施例还提供一种计算机存储介质,所述计算机可读存储介质存储有计算机程序,所述计算机程序被处理器执行以实现如上述方法实施例中记载的任何一种文本标注方法的部分或全部步骤。

[0159] 本申请实施例还提供一种计算机程序产品,所述计算机程序产品包括存储了计算机程序的非瞬时性计算机可读存储介质,所述计算机程序可操作来使计算机执行如上述方法实施例中记载的任何一种文本标注方法的部分或全部步骤。

[0160] 需要说明的是,对于前述的各方法实施例,为了简单描述,故将其都表述为一系列的动作组合,但是本领域技术人员应该知悉,本申请并不受所描述的动作顺序的限制,因为依据本申请,某些步骤可以采用其他顺序或者同时进行。其次,本领域技术人员也应该知悉,说明书中所描述的实施例均属于可选实施例,所涉及的动作和模块并不一定是本申请

所必须的。

[0161] 在上述实施例中,对各个实施例的描述都各有侧重,某个实施例中沒有详述的部分,可以参见其他实施例的相关描述。

[0162] 在本申请所提供的几个实施例中,应该理解到,所揭露的装置,可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,可以是电性或其它的形式。

[0163] 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。

[0164] 另外,在本申请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,也可以采用软件程序模块的形式实现。

[0165] 所述集成的单元如果以软件程序模块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器中。基于这样的理解,本申请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器中,包括若干指令用以使得一台计算机设备(可为个人计算机、服务器或者网络设备)执行本申请各个实施例所述方法的全部或部分步骤。而前述的存储器包括:U盘、只读存储器(ROM,Read-Only Memory)、随机存取存储器(RAM,Random Access Memory)、移动硬盘、磁碟或者光盘等各种可以存储程序代码的介质。

[0166] 本领域普通技术人员可以理解上述实施例的各种方法中的全部或部分步骤是可以通过程序来指令相关的硬件来完成,该程序可以存储于一计算机可读取存储器中,存储器可以包括:闪存盘、只读存储器(英文:Read-Only Memory,简称:ROM)、随机存取器(英文:Random Access Memory,简称:RAM)、磁盘或光盘等。

[0167] 以上对本申请实施例进行了详细介绍,本文中应用了具体个例对本申请的原理及实施方式进行了阐述,以上实施例的说明只是用于帮助理解本申请的方法及其核心思想;同时,对于本领域的一般技术人员,依据本申请的思想,在具体实施方式及应用范围上均会有改变之处,综上所述,本说明书内容不应理解为对本申请的限制。

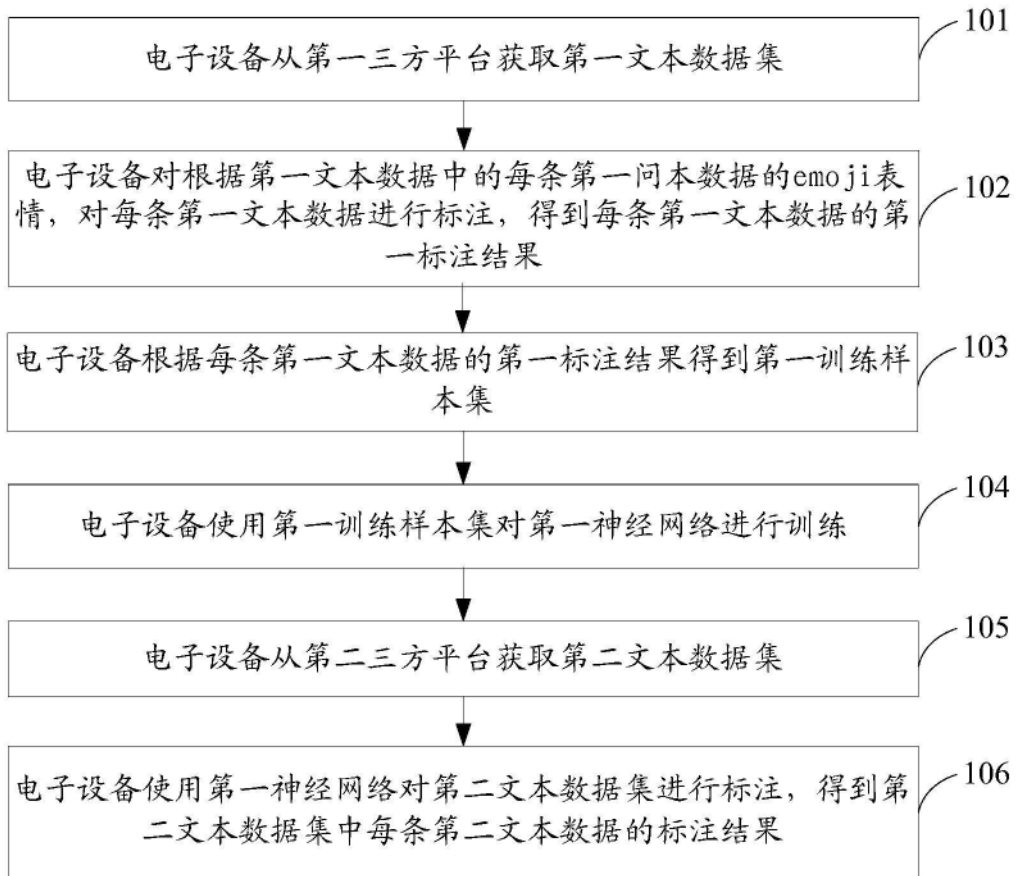


图1

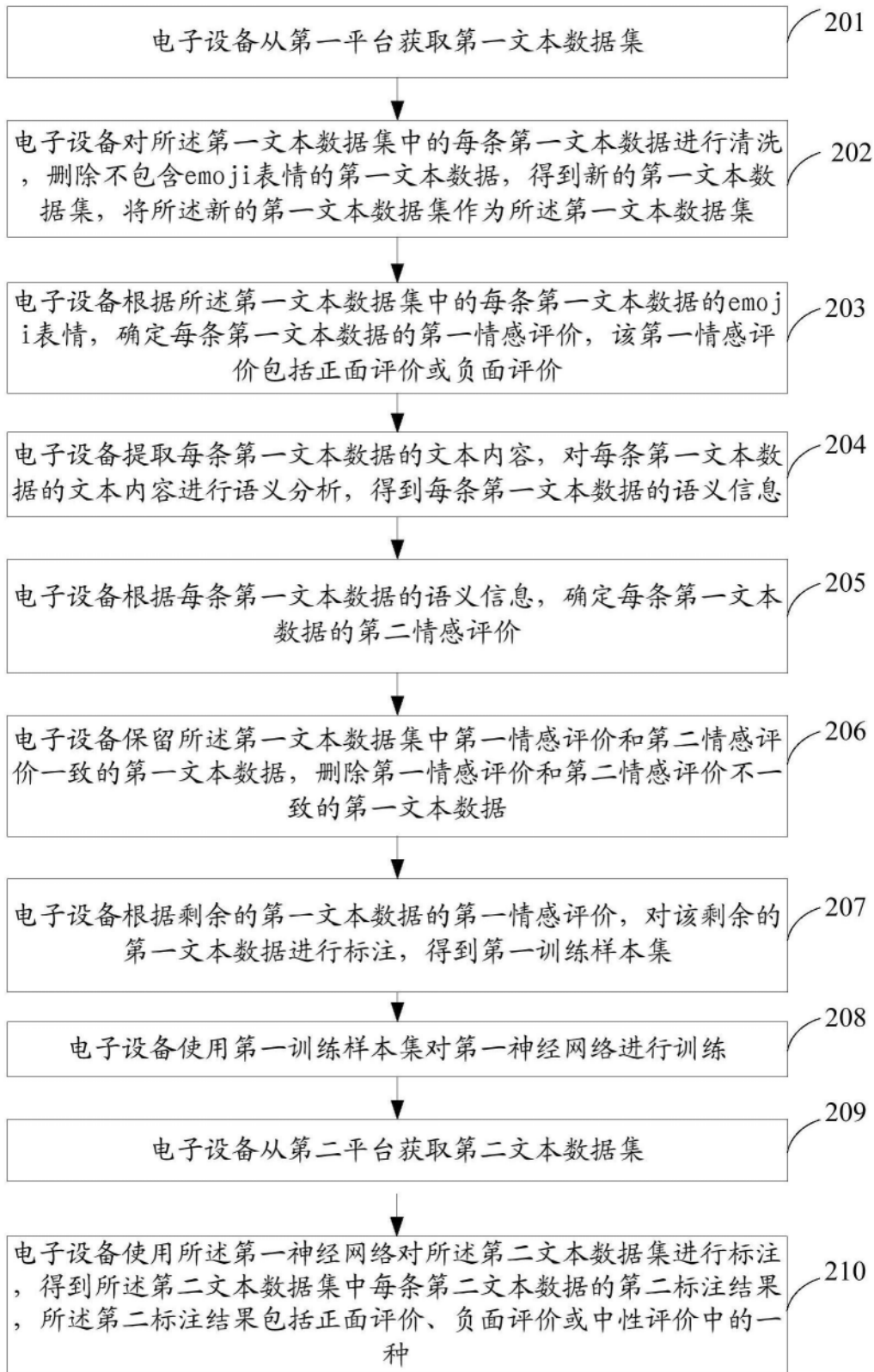


图2

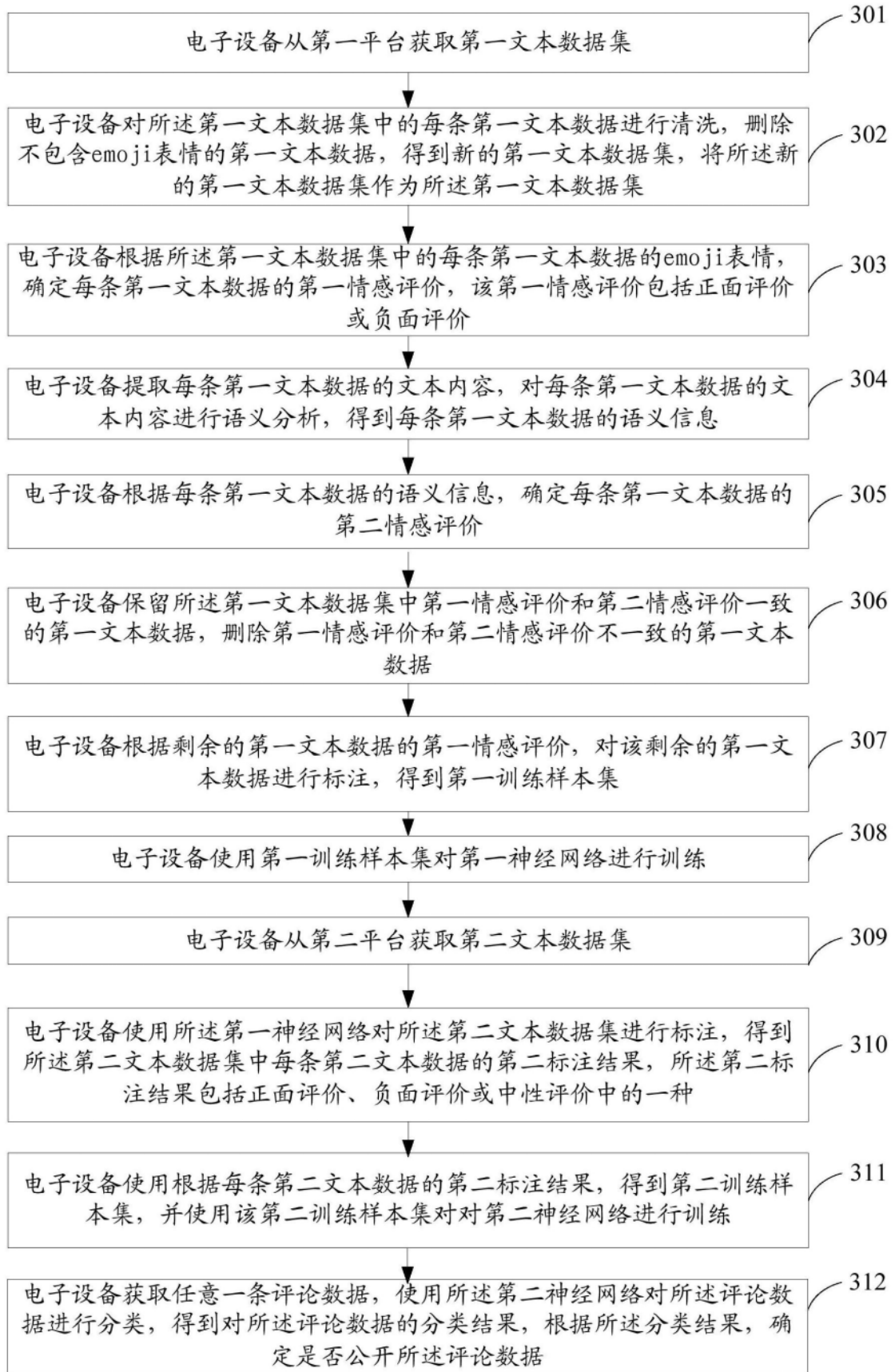


图3

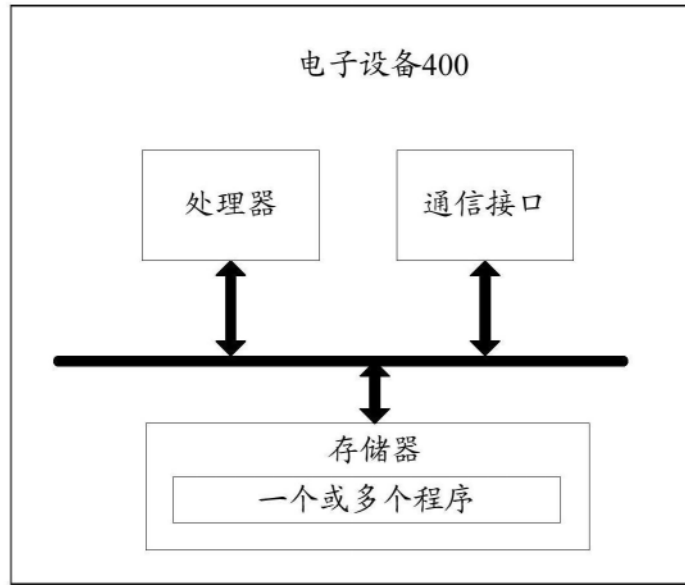


图4

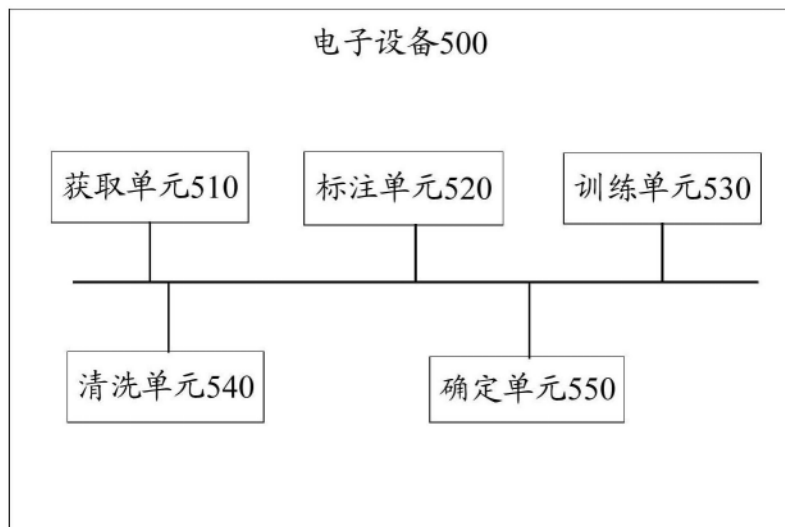


图5