(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 112214576 A
(43)申请公布日 2021.01.12

(21)申请号 202010947333.0

(22)申请日 2020.09.10

(71)申请人 深圳价值在线信息科技股份有限公司

地址 518000 广东省深圳市福田区沙头街道滨河大道9289号京基滨河时代广场北区一期B座09层

(72)发明人 赵洋 陈龙 王宇 魏世胜

(74)专利代理机构 深圳中一联合知识产权代理有限公司 44414

代理人 左婷兰

(51)Int.Cl.

G06F 16/33(2019.01)
G06F 40/211(2020.01)
G06F 40/279(2020.01)

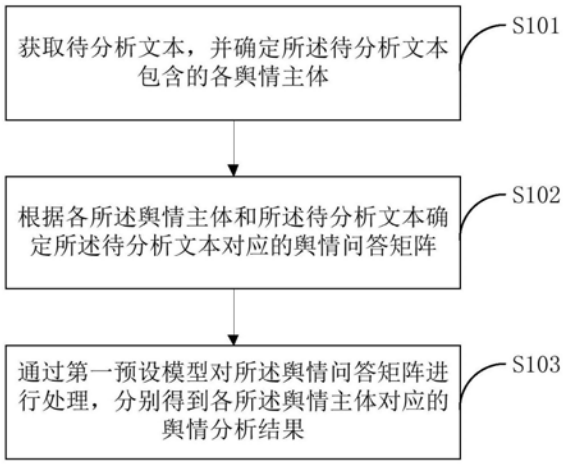
权利要求书2页 说明书10页 附图2页

(54)发明名称

舆情分析方法、装置、终端设备及计算机可读存储介质

(57)摘要

本申请适用于计算机技术领域，提供了舆情分析方法、装置、终端设备及计算机可读存储介质。所述方法包括：获取待分析文本，并确定所述待分析文本包含的各舆情主体；根据各所述舆情主体和所述待分析文本确定所述待分析文本对应的舆情问答矩阵；通过第一预设模型对所述舆情问答矩阵进行处理，分别得到各所述舆情主体对应的舆情分析结果。通过确定待分析文本中的各舆情主体，并通过第一预设模型对各舆情主体以及该舆情主体对应的语句分别进行舆情分析，可以准确得到多个舆情主体中各舆情主体对应的舆情分析结果，可有效提高包含多个舆情主体的舆情分析的准确率。



CN 112214576 A

1.一种舆情分析方法,其特征在于,包括:

获取待分析文本,并确定所述待分析文本包含的各舆情主体;

根据各所述舆情主体和所述待分析文本确定所述待分析文本对应的舆情问答矩阵,所述舆情问答矩阵包含多个舆情问答向量,每一个舆情问答向量为一个舆情主体以及所述待分析文本中与该舆情主体相关联的语句组成的向量;

通过第一预设模型对所述舆情问答矩阵进行处理,分别得到各所述舆情主体对应的舆情分析结果。

2.如权利要求1所述的舆情分析方法,其特征在于,所述根据各所述舆情主体和所述待分析文本确定所述待分析文本对应的舆情问答矩阵,包括:

获取所述待分析文本中包含舆情主体的语句;

根据所述语句和所述待分析文本,确定候选舆情语句;

根据各所述舆情主体和所述候选舆情语句得到所述舆情问答矩阵。

3.如权利要求2所述的舆情分析方法,其特征在于,所述根据所述语句和所述待分析文本,确定候选舆情语句,包括:

对所述语句和所述待分析文本中的标题进行聚类分析,得到与所述标题对应的相关语句;

获取所述待分析文本中与所述相关语句对应的上下文语句;

根据所述相关语句和所述上下文语句得到所述候选舆情语句。

4.如权利要求2所述的舆情分析方法,其特征在于,所述根据各所述舆情主体和所述候选舆情语句得到所述舆情问答矩阵,包括:

获取各所述舆情主体对应的候选舆情语句;

对各所述舆情主体和各所述舆情主体对应的候选舆情语句分别进行向量化处理,得到各所述舆情主体对应的舆情问答向量;

根据各所述舆情主体对应的舆情问答向量,得到所述舆情问答矩阵。

5.如权利要求1至4中任一项所述的舆情分析方法,其特征在于,所述确定所述待分析文本包含的各舆情主体,包括:

对所述待分析文本进行主体识别,得到所述待分析文本包含的各舆情主体。

6.如权利要求5所述的舆情分析方法,其特征在于,所述对所述待分析文本进行主体识别,得到所述待分析文本包含的各舆情主体,包括:

对所述待分析文本进行主体识别,得到所述待分析文本包含的各初始舆情主体;

获取包含相同关键词的初始舆情主体,并从所获取的初始舆情主体中确定目标舆情主体,将所述目标舆情主体确定为所述待分析文本包含的舆情主体,所述目标舆情主体为具有最大长度的初始舆情主体。

7.如权利要求1至4中任一项所述的舆情分析方法,其特征在于,所述第一预设模型为利用训练舆情文本和预设舆情分析结果训练得到的表征量模型BERT。

8.一种舆情分析装置,其特征在于,包括:

主体确定模块,用于获取待分析文本,并确定所述待分析文本包含的各实体名称舆情主体;

矩阵确定模块,用于根据各所述舆情主体和所述待分析文本确定所述待分析文本对应

的舆情问答矩阵，所述舆情问答矩阵包含多个舆情问答向量，每一个舆情问答向量为一个舆情主体以及所述待分析文本中与该舆情主体相关联的语句组成的向量；

结果分析模块，用于通过第一预设模型对所述舆情问答矩阵进行处理，分别得到各所述舆情主体对应的舆情分析结果。

9.一种终端设备，包括存储器、处理器以及存储在所述存储器中并可在所述处理器上运行的计算机程序，其特征在于，所述处理器执行所述计算机程序时实现如权利要求1至7任一项所述的舆情分析方法。

10.一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，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至7任一项所述的舆情分析方法。

舆情分析方法、装置、终端设备及计算机可读存储介质

技术领域

[0001] 本申请属于计算机技术领域，尤其涉及一种舆情分析方法、装置、终端设备及计算机可读存储介质。

背景技术

[0002] 目前，对舆情分析大部分是针对文本进行的，是判断给定的文本的情感倾向。但是，对一个文本中包含多个舆情主体进行舆情分析时，有可能会因为一个文本中包含有多个舆情主体导致舆情分析的结果有偏差；而且也不能准确的得到每个舆情主体对应的舆情分析结果。

发明内容

[0003] 本申请实施例提供了舆情分析方法、装置、终端设备及计算机可读存储介质，可以解决如何准确得到文本中包含有多个舆情主体的舆情分析结果的技术问题。

[0004] 第一方面，本申请实施例提供了一种舆情分析方法，包括：

[0005] 获取待分析文本，并确定所述待分析文本包含的各舆情主体；

[0006] 根据各所述舆情主体和所述待分析文本确定所述待分析文本对应的舆情问答矩阵，所述舆情问答矩阵包含多个舆情问答向量，每一个舆情问答向量为一个舆情主体以及所述待分析文本中与该舆情主体相关联的语句组成的向量；

[0007] 通过第一预设模型对所述舆情问答矩阵进行处理，分别得到各所述舆情主体对应的舆情分析结果。

[0008] 可选的，所述根据各所述舆情主体和所述待分析文本确定所述待分析文本对应的舆情问答矩阵，包括：

[0009] 获取所述待分析文本中包含舆情主体的语句；

[0010] 根据所述语句以及所述待分析文本，确定候选舆情语句；

[0011] 根据各所述舆情主体和所述候选舆情语句得到所述舆情问答矩阵。

[0012] 可选的，所述根据所述语句以及所述待分析文本，确定候选舆情语句，包括：

[0013] 对所述语句和所述待分析文本中的标题进行聚类分析，得到与所述标题对应的相关语句；

[0014] 获取所述待分析文本中与所述相关语句对应的各上下文语句；

[0015] 根据所述相关语句和所述上下文语句得到所述候选舆情语句。

[0016] 可选的，所述根据各所述舆情主体和所述候选舆情语句得到所述舆情问答矩阵，包括：

[0017] 获取各所述舆情主体对应的候选舆情语句；

[0018] 对各所述舆情主体和各所述舆情主体对应的候选舆情语句分别进行向量化处理，得到各所述舆情主体对应的舆情问答向量；

[0019] 根据各所述舆情主体对应的舆情问答向量，得到所述舆情问答矩阵。

[0020] 可选的，所述确定所述待分析文本包含的各舆情主体，包括：

[0021] 对所述待分析文本进行主体识别，得到所述待分析文本包含的各舆情主体。

[0022] 可选的，所述对所述待分析文本进行主体识别，得到所述待分析文本包含的各舆情主体，包括：

[0023] 对所述待分析文本进行主体识别，得到所述待分析文本包含的各初始舆情主体；

[0024] 获取包含相同关键词的初始舆情主体，并从所获取的初始舆情主体中确定目标舆情主体，将所述目标舆情主体确定为所述待分析文本包含的舆情主体，所述目标舆情主体为具有最大长度的初始舆情主体。

[0025] 可选的，所述第一预设模型为利用训练舆情文本和预设舆情分析结果训练得到的表征量模型BERT。

[0026] 第二方面，本申请实施例提供了一种舆情分析装置，包括：

[0027] 主体确定模块，用于获取待分析文本，并确定所述待分析文本包含的各实体名称舆情主体；

[0028] 矩阵确定模块，用于根据各所述舆情主体和所述待分析文本确定所述待分析文本对应的舆情问答矩阵，所述舆情问答矩阵包含多个舆情问答向量，每一个舆情问答向量为一个舆情主体以及所述待分析文本中与该舆情主体相关联的语句组成的向量；

[0029] 结果分析模块，用于通过第一预设模型对所述舆情问答矩阵进行处理，分别得到各所述舆情主体对应的舆情分析结果。

[0030] 可选的，该矩阵确定模块包括：

[0031] 语句确定单元，用于获取所述待分析文本中包含舆情主体的语句；

[0032] 候选舆情语句确定单元，用于根据所述语句以及所述待分析文本，确定候选舆情语句；

[0033] 矩阵确定单元，用于根据各所述舆情主体和所述候选舆情语句得到所述舆情问答矩阵。

[0034] 可选的，所述候选舆情语句确定单元，包括：

[0035] 聚类分析分单元，用于对所述语句和所述待分析文本中的标题进行聚类分析，得到与所述标题对应的相关语句；

[0036] 上下语句获取分单元，用于获取所述待分析文本中与所述相关语句对应的各上下文语句；

[0037] 候选舆情语句分单元，用于根据所述相关语句和所述上下文语句得到所述候选舆情语句。

[0038] 可选的，所述矩阵确定单元，包括：

[0039] 主体对应语句确定分单元，用于获取各所述舆情主体对应的候选舆情语句；

[0040] 向量确定分单元，用于对各所述舆情主体和各所述舆情主体对应的候选舆情语句分别进行向量化处理，得到各所述舆情主体对应的舆情问答向量；

[0041] 矩阵确定分单元，用于根据各所述舆情主体对应的舆情问答向量，得到所述舆情问答矩阵。

[0042] 可选的，所述主体确定模块，包括：

[0043] 主体确定单元，用于对所述待分析文本进行主体识别，得到所述待分析文本包含

的各舆情主体。

[0044] 可选地，主体确定单元，包括：

[0045] 初始舆情主体确定分单元，用于对所述待分析文本进行主体识别，得到所述待分析文本包含的各初始舆情主体；

[0046] 舆情主体确定分单元，用于获取包含相同关键词的初始舆情主体，并从所获取的初始舆情主体中确定目标舆情主体，将所述目标舆情主体确定为所述待分析文本包含的舆情主体，所述目标舆情主体为具有最大长度的初始舆情主体。

[0047] 第三方面，本申请实施例提供了一种终端设备，包括存储器、处理器以及存储在所述存储器中并可在所述处理器上运行的计算机程序，所述处理器执行所述计算机程序时实现如上述第一方面中任一项所述的舆情分析方法。

[0048] 第四方面，本申请实施例提供了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，所述计算机程序被处理器执行时实现如上述第一方面中任一项所述的舆情分析方法。

[0049] 第五方面，本申请实施例提供了一种计算机程序产品，当计算机程序产品在终端设备上运行时，使得终端设备执行上述第一方面中任一项所述的舆情分析方法。

[0050] 本申请实施例与现有技术相比存在的有益效果是：

[0051] 本申请实施例中，通过确定待分析文本包含的各舆情主体，并确定待分析文本中与各舆情主体相关联的语句；然后对舆情主体以及该舆情主体相关联的语句进行向量化处理，得到各舆情主体对应的舆情问答向量，并将各舆情问答向量组成舆情问答矩阵输入至第一预设模型进行处理，从而分别得到各舆情主体对应的舆情分析结果，即通过确定待分析文本中的各舆情主体，并通过第一预设模型对各舆情主体以及该舆情主体对应的语句分别进行舆情分析，可以准确得到多个舆情主体中各舆情主体对应的舆情分析结果，可有效提高包含多个舆情主体的舆情分析的准确率。

附图说明

[0052] 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的前提下，还可以根据这些附图获得其他的附图。

[0053] 图1是本申请实施例一提供的舆情分析方法的流程示意图；

[0054] 图2是本申请实施例二提供的舆情分析方法的流程示意图；

[0055] 图3是本申请实施例提供的舆情分析装置的结构示意图；

[0056] 图4是本申请实施例提供的终端设备的结构示意图。

具体实施方式

[0057] 以下描述中，为了说明而不是为了限定，提出了诸如特定系统结构、技术之类的具体细节，以便透彻理解本申请实施例。然而，本领域的技术人员应当清楚，在没有这些具体细节的其它实施例中也可以实现本申请。在其它情况中，省略对众所周知的系统、装置、电路以及方法的详细说明，以免不必要的细节妨碍本申请的描述。

[0058] 应当理解，当在本申请说明书和所附权利要求书中使用时，术语“包括”指示所描述特征、整体、步骤、操作、元素和/或组件的存在，但并不排除一个或多个其它特征、整体、步骤、操作、元素、组件和/或其集合的存在或添加。

[0059] 还应当理解，在本申请说明书和所附权利要求书中使用的术语“和/或”是指相关联列出的项中的一个或多个的任何组合以及所有可能组合，并且包括这些组合。

[0060] 如在本申请说明书和所附权利要求书中所使用的那样，术语“如果”可以依据上下文被解释为“当...时”或“一旦”或“响应于确定”或“响应于检测到”。类似地，短语“如果确定”或“如果检测到[所描述条件或事件]”可以依据上下文被解释为意指“一旦确定”或“响应于确定”或“一旦检测到[所描述条件或事件]”或“响应于检测到[所描述条件或事件]”。

[0061] 另外，在本申请说明书和所附权利要求书的描述中，术语“第一”、“第二”、“第三”等仅用于区分描述，而不能理解为指示或暗示相对重要性。

[0062] 在本申请说明书中描述的参考“一个实施例”或“一些实施例”等意味着在本申请的一个或多个实施例中包括结合该实施例描述的特定特征、结构或特点。由此，在本说明书中的不同之处出现的语句“在一个实施例中”、“在一些实施例中”、“在其他一些实施例中”、“在另外一些实施例中”等不是必然都参考相同的实施例，而是意味着“一个或多个但不是所有的实施例”，除非是以其他方式另外特别强调。术语“包括”、“包含”、“具有”及它们的变形都意味着“包括但不限于”，除非是以其他方式另外特别强调。

[0063] 本申请实施例提供的舆情分析方法可以应用于手机、平板电脑、可穿戴设备、车载设备、增强现实(augmented reality,AR)/虚拟现实(virtual reality,VR)设备、笔记本电脑、超级移动个人计算机(ultra-mobile personal computer,UMPC)、上网本、个人数字助理(personal digital assistant,PDA)等终端设备上，本申请实施例对终端设备的具体类型不作任何限制。

[0064] 本申请实施例提供的舆情分析方法的执行主体为终端设备，终端设备获取待分析文本，并确定所述待分析文本包含的各舆情主体；终端设备根据各所述舆情主体和所述待分析文本确定所述待分析文本对应的舆情问答矩阵，所述舆情问答矩阵包含多个舆情问答向量，每一个舆情问答向量为一个舆情主体以及所述待分析文本中与该舆情主体相关联的语句组成的向量；终端设备通过第一预设模型对所述舆情问答矩阵进行处理，分别得到各所述舆情主体对应的舆情分析结果。

[0065] 图1示出了本申请实施例提供的舆情分析方法的示意性流程图，作为示例而非限定，该方法可以应用于上述终端设备中。如图1所示，所述方法可以包括：

[0066] S101、获取待分析文本，并确定所述待分析文本包含的各舆情主体；

[0067] 其中，待分析文本是需要进行舆情分析的文本，待分析文本中可以包含有一个舆情主体，也可以包含有多个舆情主体。终端设备获取到待分析文本后，可以对待分析文本进行主体识别，得到待分析文本中所包含的各舆情主体。其中，舆情主体可以是待分析文本中包含的公司名称，也可以是待分析文本中包含的人物名称。例如，在待分析文本中包含“AA公司、管理会和BB”等舆情主体时，终端设备在获取待分析文本后，可以通过主体识别模型识别出待分析文本中的“AA公司”、“管理会”和“BB”。

[0068] 其中，主体识别模型可以是长短期记忆网络(Long Short-Term Memory,LSTM)与条件随机场(conditional random field,CRF)构成的模型，也可以是空洞卷积(iterated

dilated convolutional neural network,ID-CNN)与条件随机场(conditional random field,CRF)构成的模型。

[0069] 在一种可能的实现方式中,对所述待分析文本进行主体识别,得到所述待分析文本包含的各舆情主体可以是:对所述待分析文本进行主体识别,得到所述待分析文本包含的各初始舆情主体;获取包含相同关键词的初始舆情主体,并从所获取的初始舆情主体中确定目标舆情主体,将所述目标舆情主体确定为所述待分析文本包含的舆情主体,所述目标舆情主体为具有最大长度的初始舆情主体。

[0070] 在该实现方式中,终端设备在获取了待分析文本之后,可以通过主体识别模型对待分析文本进行主体识别,得到待分析文本包含的若干个初始舆情主体。然后,终端设备再对若干个初始舆情主体进行筛选,以获取包含相同关键词的初始舆情主体,并从所获取的初始舆情主体中确定目标舆情主体;然后终端设备可以将目标舆情主体确定为所述待分析文本包含的舆情主体,所述目标舆情主体为具有最大长度的初始舆情主体。

[0071] 例如,在待分析文本中包含“AA公司、AA、美国AA公司、管理会、美国管理会、BB公司和BB”等舆情主体时,终端设备获取到待分析文本后,可以通过主体识别模型识别出待分析文本中的“AA公司、AA、美国AA公司、管理会、美国管理会、BB公司和BB”等初始舆情主体。然后,终端设备再各初始舆情主体进行筛选,可以获取到具有相同的关键词“AA”的“AA公司”、“AA”和“美国AA公司”等初始舆情主体,并可以从“AA公司”、“AA”和“美国AA公司”等初始舆情主体中确定出目标舆情主体为“美国AA公司”。同时,还可以获取到具有相同的关键词“管理会”的“管理会”和“美国管理会”等初始舆情主体,并可以从“管理会”和“美国管理会”等初始舆情主体中确定出目标舆情主体为“美国管理会”。同时,还可以获取到具有相同的关键词“BB”的“BB公司”和“BB”等初始舆情主体,并可以从“BB公司”和“BB”等初始舆情主体中确定出目标舆情主体为“BB公司”。然后,终端设备将“美国AA公司”、“美国管理会”和“BB公司”确定为待分析文本中所包含的舆情主体。

[0072] S102、根据各所述舆情主体和所述待分析文本确定所述待分析文本对应的舆情问答矩阵,所述舆情问答矩阵包含多个舆情问答向量,每一个舆情问答向量为一个舆情主体以及所述待分析文本中与该舆情主体相关联的语句组成的向量;

[0073] 在本实施例中,终端设备在得到待分析文本包含的各舆情主体后,可以在待分析文本中分别确定与各舆情主体相关联的语句,并将一个舆情主体与该舆情主体相关联的语句组成舆情问答向量;再将若干个舆情主体对应的舆情问答向量组成舆情问答矩阵。其中,语句为待分析文本中的一个自然句。舆情主体相关联的语句可以是包含舆情主体的语句,也可以是包含舆情主体的语句、该语句的上一个或上多个语句以及该语句的下一个或下多个语句。舆情问答向量中与舆情主体相关联的语句可以是一个句子,也可以是多个句子。

[0074] 例如,一个舆情主体为“美国AA公司”时,终端设备在待分析文本中确定与“美国AA公司”相关的句子“美国AA公司麻烦缠身,被CC起诉侵权”,终端设备将“美国AA公司”和“美国AA公司麻烦缠身,被CC起诉侵权”组成舆情问答向量。然后,终端设备将若干个舆情问答向量组成舆情问答矩阵。

[0075] 在一种可能的实现方式中,如图2所示,S102可以包括S201、S202和S203。

[0076] S201、获取所述待分析文本中包含舆情主体的语句;

[0077] 具体的,终端设备在确定待分析文本包含的各舆情主体后,可以在待分析文本筛

选出包含舆情主体的所有语句。每一个语句可以为待分析文本中的一个自然句。

[0078] S202、根据所述语句以及所述待分析文本，确定候选舆情语句；

[0079] 具体的，终端设备在获取包含舆情主体的所有语句之后，在待分析文本中获取与各个语句相关联的候选舆情语句。候选舆情语句可以是包含有舆情主体的语句，也可以是包含舆情主体的语句、该语句的上一个或上多个语句以及该语句的下一个或下多个语句。候选舆情语句可以是一个语句，也可以是多个语句。

[0080] 在一种可能的实现方式中，根据所述语句以及所述待分析文本，确定候选舆情语句可以是：对所述语句和所述待分析文本中的标题进行聚类分析，得到与所述标题对应的相关语句；获取所述待分析文本中与所述相关语句对应的各上下文语句；根据所述相关语句和所述上下文语句得到所述候选舆情语句。

[0081] 在该实现方式中，终端设备在获取包含舆情主体的所有语句后，可以将包含舆情主体的所有语句和待分析文本中的标题通过k均值聚类算法(k-means clustering algorithm)进行聚类分析。具体地，将包含舆情主体的所有语句和待分析文本中的标题分为2组，随机选取2个对象作为初始的聚类中心，然后计算每个对象与各个聚类中心之间的距离，把每个对象分配给距离它最近的聚类中心。聚类中心以及分配给它们的对象就代表一个类簇。在完成所有对象的一次分配后，终端设备可以根据类簇中现有的对象重新计算聚类中心，然后重新执行聚类操作。这个过程将不断重复直到满足某个终止条件。当满足终止条件时，聚类终止，终端设备可以将标题所在的类簇中的所有语句确定为与标题对应的相关语句。

[0082] 其中，终止条件可以是各类簇中对象的变化数量小于或等于预设数目，或者可以是小于预设数目的聚类中心再发生变化，或者可以是误差平方和局部最小。对象可以为向量化处理后的一个语句或标题。

[0083] 在该实现方式中，终端设备得到所有的相关语句后，获取待分析文本中与各个相关语句分别对应的上下文语句。终端设备分别将各个相关语句和与该相关语句对应的上下文语句依次组合，得到各个相关语句对应的候选舆情语句。然后将所有的候选舆情语句组合在一起形成候选舆情语句。其中，上下文语句为在待分析文本排在相关语句的前一句或前几句的语句以及在待分析文本排在相关语句的后一句或后几句的语句。

[0084] S203、根据各所述舆情主体和所述候选舆情语句得到所述舆情问答矩阵。

[0085] 具体的，终端设备在确定出候选舆情语句之后，将一个舆情主体以及与该舆情主体对应的候选舆情语句组合生成舆情问答向量；再将若干个舆情主体对应的舆情问答向量组成舆情问答矩阵。

[0086] 在一种可能的实现方式中，根据各所述舆情主体和所述候选舆情语句得到所述舆情问答矩阵可以是：获取各所述舆情主体对应的候选舆情语句；对各所述舆情主体和各所述舆情主体对应的候选舆情语句分别进行向量化处理，得到各所述舆情主体对应的舆情问答向量；根据各所述舆情主体对应的舆情问答向量，得到所述舆情问答矩阵。

[0087] 在该实现方式中，终端设备在获取到所有的候选舆情语句之后，可以分别获取各舆情主体对应的候选舆情语句。然后，可以通过分词工具对各个候选舆情语句进行分词，并通过针对关键词的统计分析方法(term frequency-inverse document frequency，TF-IDF)对各个候选舆情语句进行向量化处理。向量化处理后的候选舆情语句可以表示为向量

$[F_1,F_2,...,F_i,...,F_{W_n}]$,$F_i=TFIDF_i$。其中，$TFIDF_i=\log(\frac{tf_i}{\sum_{j=1}^{|D|}1\{i\in d_j\}+\varepsilon})$，$D=\{d_1,d_2,...,d_n\}$，$tf_i=\frac{freq_i}{\sum_{j=1}^{W_n}freq_j}$，$tf_i$为候选舆情语句中第i个词的词频，D为训练集文本数，W_n为词表大小，ε为平滑因子，$1\{i\in d_j\}$表示如果d_j包含词i，则为1。然后，可以通过TF-IDF对各个舆情主体近向量化处理。然后，可以将完成向量化的各个舆情主体以及与舆情主体对应且完成向量化的候选舆情语句组合，得到各个舆情主体对应的舆情问答向量。然后，终端设备按照舆情主体的顺序依次将舆情问答向量组合在一起，形成舆情问答矩阵。其中，一个舆情主体对应的候选舆情语句可以包含有一个或者多个语句。舆情问答向量的格式可以为[该舆情主体对应的候选舆情语句，舆情主体]。舆情问答矩阵为多个舆情问答向量组成的矩阵。

[0088] S103、通过第一预设模型对所述舆情问答矩阵进行处理，分别得到各所述舆情主体对应的舆情分析结果。

[0089] 在本申请实施例中，终端设备得到舆情问答矩阵后，将舆情问答矩阵中的每个舆情问答向量中的舆情主体和与该舆情主体相关联的候选舆情语句分别以句2和句1输入第一预设模型进行处理，分别得到各舆情主体对应的舆情分析结果。其中，第一预设模型可以是利用训练舆情文本和预设舆情分析结果训练得到的表征量模型(Bidirectional Encoder Representations from Transformers，BERT)。训练舆情文本为对BERT模型进行训练的文本，预设舆情分析结果为训练舆情文本中，每个训练舆情主体对应的舆情分析结果。预设舆情分析结果为给定的值。预设舆情分析结果可以为正面、负面、中性。

[0090] 在对第一预设模型进行训练时，先通过与S101和S102相同的方式对训练舆情文本进行处理，得到训练舆情问答矩阵。再将训练舆情问答矩阵中的每个训练舆情问答向量中的训练舆情主体和与该训练舆情主体相关联的训练语句分别以句2和句1方式输入BERT模型中。然后，再将预设舆情分析结果输入BERT模型。通过训练舆情主体、与该训练舆情主体相关联的训练语句以及预设舆情分析结果对BERT模型进行训练得到第一预设模型。其中，训练舆情主体为训练舆情文本包含中的舆情主体，训练舆情问答矩阵为通过S101和S102的方式处理训练舆情文本得到的舆情问答矩阵。

[0091] 具体的，在对第一预设模型进行训练时，终端设备先通过与S101和S102相同的方式对训练舆情文本进行处理，得到各训练舆情主体和各训练舆情主体对应的训练舆情语句，并标记各训练舆情主体对应的预设舆情分析结果。然后，终端设备可以对每一个训练舆情主体以及该训练舆情主体对应的训练舆情语句进行向量化处理，从而得到各训练舆情主体对应的舆情问答向量$[seg_i,ent_i]$。然后，终端设备可以将舆情问答向量$[seg_i,ent_i]$和该舆情问答向量对应的预设舆情分析结果$label_i$，输入BERT模型进行处理，得到各训练舆情主体对应的训练舆情分析结果。其中，ent_i为第i个训练舆情主体，seg_i为第i个训练舆情主体对应的训练舆情语句，$label_i$为第i个训练舆情主体对应的预设舆情分析结果。随后，终端设备可以根据训练舆情分析结果与预设舆情分析结果计算训练误差。当训练误差大于预设误差时，终端设备可以调整BERT模型的模型参数，并利用训练舆情文本和所对应的预设舆情分析结果对调整后的BERT模型继续进行训练，直至训练误差小于或等于预设误差为

止，从而得到训练后的BERT模型。

[0092] 具体的，终端设备通过主体识别模型识别待分析文本中的舆情主体时，得到的舆情主体可以表示为$\{ent_i, index_i^{begin}, index_i^{end}\}$。其中，$ent_i$为第i个舆情主体，$index_i^{begin}$和$index_i^{end}$分别为该舆情主体在待分析文本中的坐标。然后，终端设备根据舆情主体的坐标，确定包含舆情主体的语句，包含舆情主体的语句可以表示为$S_i=[s_1,s_2,...,s_{N_i}]$。其中，一个舆情主体可能在文本的多个句子中出现，$N_i$为舆情主体i在待分析文本中出现的句子个数。然后，终端设备对所有的语句S_i和待分析文本中的标题s_T进行聚类分析，得到与标题s_T对应的相关语句$S_c=[s_1,s_2,...,s_j]$，其中s_j为相关语句中的一个句子。然后，终端设备获取待分析文本中与该相关语句s_j对应的上下文语句s_{j-n},s_{j+n}。其中，n为自然数。然后，终端设备根据一个相关语句s_j和该相关语句的上下文语句s_{j-n},s_{j+n}得到一个候选舆情语句$c_j=[s_{j-n},s_j,s_{j+n}]$。然后，终端设备获取各舆情主体i对应的候选舆情语句$C_i=[c_1,c_2,...,c_{|S_c|}]$，$C_i$为与舆情主体对应的候选舆情语句。然后，终端设备对各所述舆情主体ent_i和各所述舆情主体对应的候选舆情语句C_i分别进行向量化处理，得到各所述舆情主体对应的舆情问答向量$[C_i,ent_i]$。然后，终端设备将舆情主体对应的候选舆情语句C_i确定为句2和舆情主体ent_i确定为句1的方式输入BERT模型进行处理，得到舆情主体ent_i对应的舆情分析结果。BERT模型在接收到C_i和ent_i之后，可以将C_i和ent_i中的每个词转换为词嵌入(Token Embedding)、片段嵌入(Segment Embedding)和位置嵌入(Position Embedding)的加和，并以CLS为开始标志，SEP为两个句子的分隔符。然后，BERT模型对转换后的C_i和ent_i进行舆情分析，得到舆情分析结果。

[0093] 综上所述，通过确定待分析文本包含的各舆情主体，并确定待分析文本中与各舆情主体相关联的语句；然后对舆情主体以及该舆情主体相关联的语句进行向量化处理，得到各舆情主体对应的舆情问答向量，并将各舆情问答向量组成舆情问答矩阵输入至第一预设模型进行处理，从而分别得到各舆情主体对应的舆情分析结果，即通过确定待分析文本中的各舆情主体，并通过第一预设模型对各舆情主体以及该舆情主体对应的语句分别进行舆情分析，可以准确得到多个舆情主体中各舆情主体对应的舆情分析结果，可有效提高包含多个舆情主体的舆情分析的准确率。

[0094] 对应于上文实施例所述的舆情分析方法，图3示出了本申请实施例提供的舆情分析装置的结构框图，为了便于说明，仅示出了与本申请实施例相关的部分。

[0095] 参照图3，该装置包括：

[0096] 主体确定模块301，用于获取待分析文本，并确定所述待分析文本包含的各实体名称舆情主体；

[0097] 矩阵确定模块302，用于根据各所述舆情主体和所述待分析文本确定所述待分析文本对应的舆情问答矩阵，所述舆情问答矩阵包含多个舆情问答向量，每一个舆情问答向量为一个舆情主体以及所述待分析文本中与该舆情主体相关联的语句组成的向量；

[0098] 结果分析模块303，用于通过第一预设模型对所述舆情问答矩阵进行处理，分别得到各所述舆情主体对应的舆情分析结果。

[0099] 可选的，该确定模块302可以包括：

[0100] 语句确定单元，用于获取所述待分析文本中包含舆情主体的语句；

[0101] 候选舆情语句确定单元，用于根据所述语句以及所述待分析文本，确定候选舆情语句；

[0102] 矩阵确定单元，用于根据各所述舆情主体和所述候选舆情语句得到所述舆情问答矩阵。

[0103] 可选的，候选舆情语句确定单元可以包括：

[0104] 聚类分析分单元，用于对所述语句和所述待分析文本中的标题进行聚类分析，得到与所述标题对应的相关语句；

[0105] 上下语句获取分单元，用于获取所述待分析文本中与所述相关语句对应的各上下文语句；

[0106] 候选舆情语句分单元，用于根据所述相关语句和所述上下文语句得到所述候选舆情语句。

[0107] 可选的，矩阵确定单元可以包括：

[0108] 主体对应语句确定分单元，用于获取各所述舆情主体对应的候选舆情语句；

[0109] 向量确定分单元，用于对各所述舆情主体和各所述舆情主体对应的候选舆情语句分别进行向量化处理，得到各所述舆情主体对应的舆情问答向量；

[0110] 矩阵确定分单元，用于根据各所述舆情主体对应的舆情问答向量，得到所述舆情问答矩阵。

[0111] 可选的，主体确定模块301可以包括：

[0112] 主体确定单元，用于对所述待分析文本进行主体识别，得到所述待分析文本包含的各舆情主体。

[0113] 可选地，主体确定单元可以包括：

[0114] 初始舆情主体确定分单元，用于对所述待分析文本进行主体识别，得到所述待分析文本包含的各初始舆情主体；

[0115] 舆情主体确定分单元，用于获取包含相同关键词的初始舆情主体，并从所获取的初始舆情主体中确定目标舆情主体，将所述目标舆情主体确定为所述待分析文本包含的舆情主体，所述目标舆情主体为具有最大长度的初始舆情主体。

[0116] 需要说明的是，上述装置/单元之间的信息交互、执行过程等内容，由于与本申请方法实施例基于同一构思，其具体功能及带来的技术效果，具体可参见方法实施例部分，此处不再赘述。

[0117] 所属领域的技术人员可以清楚地了解到，为了描述的方便和简洁，仅以上述各功能单元、模块的划分进行举例说明，实际应用中，可以根据需要而将上述功能分配由不同的功能单元、模块完成，即将所述装置的内部结构划分成不同的功能单元或模块，以完成以上描述的全部或者部分功能。实施例中的各功能单元、模块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。另外，各功能单元、模块的具体名称也只是为了便于相互区分，并不用于限制本申请的保护范围。上述系统中单元、模块的具体工作过程，可以参考前述方法实施例中的对应过程，在此不再赘述。

[0118] 本申请实施例还提供了一种终端设备，该终端设备包括：至少一个处理器、存储器以及存储在所述存储器中并可在所述至少一个处理器上运行的计算机程序，所述处理器执

行所述计算机程序时实现上述任意各个方法实施例中的步骤。

[0119] 本申请实施例还提供了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，所述计算机程序被处理器执行时可实现上述各个方法实施例中的步骤。

[0120] 本申请实施例提供了一种计算机程序产品，当计算机程序产品在终端设备上运行时，使得终端设备执行时可实现上述各个方法实施例中的步骤。

[0121] 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时，可以存储在一个计算机可读存储介质中。基于这样的理解，本申请实现上述实施例方法中的全部或部分流程，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计算机可读存储介质中，该计算机程序在被处理器执行时，可实现上述各个方法实施例的步骤。其中，所述计算机程序包括计算机程序代码，所述计算机程序代码可以为源代码形式、对象代码形式、可执行文件或某些中间形式等。所述计算机可读存储介质至少可以包括：能够将计算机程序代码携带到装置/终端设备的任何实体或装置、记录介质、计算机存储器、只读存储器(read-only memory，ROM，)、随机存取存储器(random access memory，RAM，)、电载波信号、电信信号以及软件分发介质。例如U盘、移动硬盘、磁碟或者光盘等。在某些司法管辖区，根据立法和专利实践，计算机可读存储介质不可以是电载波信号和电信信号。

[0122] 在上述实施例中，对各个实施例的描述都各有侧重，某个实施例中没有详述或记载的部分，可以参见其它实施例的相关描述。

[0123] 本领域普通技术人员可以意识到，结合本文中所公开的实施例描述的各示例的单元及算法步骤，能够以电子硬件、或者计算机软件和电子硬件的结合来实现。这些功能究竟以硬件还是软件方式来执行，取决于技术方案的特定应用和设计约束条件。专业技术人员可以对每个特定的应用来使用不同方法来实现所描述的功能，但是这种实现不应认为超出本申请的范围。

[0124] 在本申请所提供的实施例中，应该理解到，所揭露的装置/终端设备和方法，可以通过其它的方式实现。例如，以上所描述的装置/终端设备实施例仅仅是示意性的，例如，所述模块或单元的划分，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讯连接可以是通过一些接口，装置或单元的间接耦合或通讯连接，可以是电性，机械或其它的形式。

[0125] 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。

[0126] 以上所述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，均应包含在本申请的保护范围之内。

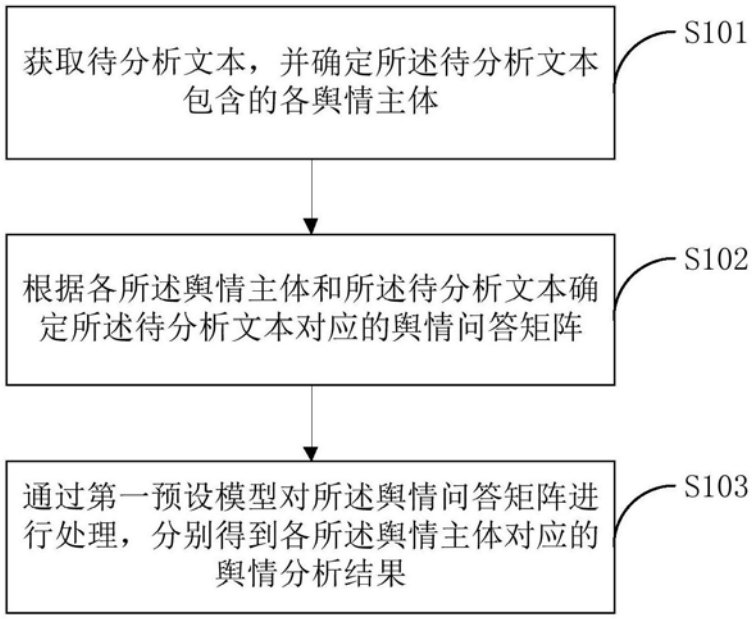

获取待分析文本，并确定所述待分析文本包含的各舆情主体
S101
根据各所述舆情主体和所述待分析文本确定所述待分析文本对应的舆情问答矩阵
S102
通过第一预设模型对所述舆情问答矩阵进行处理，分别得到各所述舆情主体对应的舆情分析结果
S103

图1

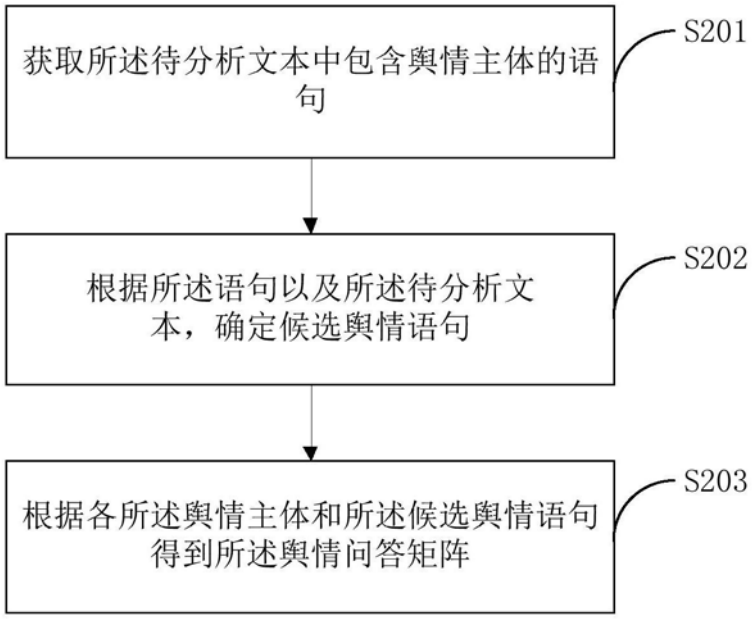

获取所述待分析文本中包含舆情主体的语句
S201
根据所述语句以及所述待分析文本，确定候选舆情语句
S202
根据各所述舆情主体和所述候选舆情语句得到所述舆情问答矩阵
S203

图2

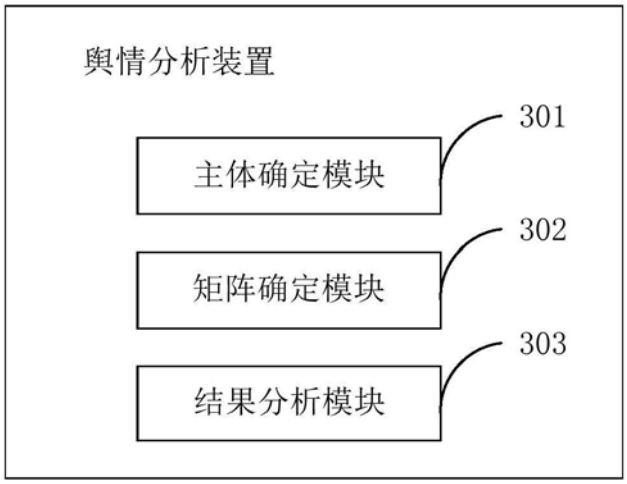

舆情分析装置
301
主体确定模块
302
矩阵确定模块
303
结果分析模块

图3

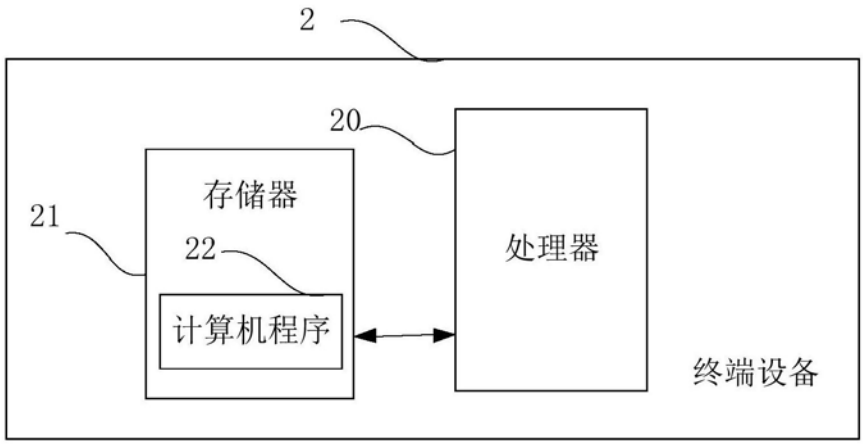

2
20
存储器
21
22
计算机程序
处理器
终端设备

图4