

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局



(43) 国际公布日
2019 年 11 月 14 日 (14.11.2019)

WIPO | PCT

(10) 国际公布号
WO 2019/214145 A1

(51) 国际专利分类号:

G06F 17/27 (2006.01) **G06K 9/62** (2006.01)
G06N 3/04 (2006.01)

福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518000 (CN)。

(21) 国际申请号: PCT/CN2018/107725

(22) 国际申请日: 2018 年 9 月 26 日 (26.09.2018)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:

201810443238.X 2018 年 5 月 10 日 (10.05.2018) CN

(74) 代理人: 深圳市沃德知识产权代理有限公司 (普通合伙) (SHENZHEN WORLD INTELLECTUAL PROPERTY AGENCY (GENERAL PARTNERSHIP)); 中国广东省深圳市福田区园岭街道八卦四路 10 号中浩大厦 1528-1530 室于志光, Guangdong 518000 (CN)。

(71) 申请人: 平安科技 (深圳) 有限公司 (PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) [CN/CN]; 中国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, Guangdong 518000 (CN)。

(81) 指定国 (除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。

(72) 发明人: 李正洋 (LI, Zhengyang); 中国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, Guangdong 518000 (CN)。李海疆 (LI, Haijiang); 中国广东省深圳市福田区

(54) Title: TEXT SENTIMENT ANALYZING METHOD, APPARATUS AND STORAGE MEDIUM

(54) 发明名称: 文本情绪分析方法、装置及存储介质

接收携带目标文本的文本情绪分析请求, 对所述目标文本进行预处理, 并采用预设序列标注法对经过预处理后的目标文本进行分词处理, 得到该目标文本对应的可用词集合

S1 S1 Receive a text sentiment analysis request carrying a target text, pre-process the target text, and perform word partition processing on the pre-processed target text using a preset sequence marking method, to obtain an available word set corresponding to the target text

确定该目标文本的待分析的句子, 根据该目标文本对应的可用词集合获取所述目标文本中待分析的句子对应的可用词集合, 根据预设计算规则, 分别计算所述待分析的句子向量的句子向量, 句子向量包括句子级别的向量和词语级别的向量

S2 S2 Determine sentences to be analyzed of the target text, acquire, according to the available word set corresponding to the target text, available word sets corresponding to said sentences in the target text, and calculate, according to a preset calculation rule, sentence vectors of said sentences, respectively, said sentence vectors comprising sentence-level vectors and word-level vectors

分别将所述待分析的句子向量的句子向量输入预先训练好的情绪判断模型, 根据模型输出结果判断所述目标文本的情绪极性

S3 S3 Input the sentence vectors of said sentences into a pre-trained sentiment determination model, and determine the sentiment nature of the target text according to an output result of the model

图 1

(57) Abstract: A text sentiment analyzing method, said method comprising: receiving a text sentiment analysis request carrying a target text, pre-processing the target text, and performing word partition processing on the pre-processed target text using a preset sequence marking method, to obtain an available word set corresponding to the target text (S1); determining sentences to be analyzed of the target text, acquiring, according to the available word set corresponding to the target text, available word sets corresponding to said sentences in the target text, and calculating, according to a preset calculation rule, sentence vectors of said sentences, respectively, said sentence vectors comprising sentence-level vectors and word-level vectors (S2); and inputting the sentence vectors of said sentences into a pre-trained sentiment determination model, and determining the sentiment nature of the target text according to an output result of the model (S3). Also provided are an electronic apparatus and a computer storage medium, being able to improve the accuracy of sentiment analysis concerning a target text.

(84) 指定国 (除另有指明, 要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告 (条约第21条(3))。

(57) 摘要: 一种文本情绪分析方法, 该方法包括: 接收携带目标文本的文本情绪分析请求, 对所述目标文本进行预处理, 并采用预设序列标注法对经过预处理后的目标文本进行分词处理, 得到该目标文本对应的可用词集合 (S1); 确定该目标文本的待分析的句子, 根据该目标文本对应的可用词集合获取所述目标文本中待分析的句子对应的可用词集合, 根据预设计算规则, 分别计算所述待分析的句子句子向量, 句子向量包括句子级别的向量和词语级别的向量 (S2); 及, 分别将所述待分析的句子句子向量输入预先训练好的情绪判断模型, 根据模型输出结果判断所述目标文本的情绪极性 (S3)。还提供一种电子装置及计算机存储介质, 可以提高目标文本情绪分析的准确性。

文本情绪分析方法、装置及存储介质

本申请要求于2018年5月10日提交中国专利局，申请号为201810443238.X、
发明名称为“文本情绪分析方法、装置及存储介质”的中国专利申请的优先
5 权，其全部内容通过引用结合在本申请中。

技术领域

本申请涉及数据处理技术领域，尤其涉及一种文本情绪分析方法、电子
装置及计算机可读存储介质。

10

背景技术

对于投资机构、投资者制定交易策略来说，特定事件或人物的相应文本
的分析是十分必要且有用的。如我们通过分析某人(例如，X先生)在twitter 上
发布的所有关于中美贸易战的言论，我们大致可以了解其对该事件的态度，
15 这对于中方提前做相应的准备是十分有用的。然而如今对于这类事件及人物
的分析主要依靠于专业人员的人工作业，此类方法主要依赖于研究人员的专业
能力及个人经验，很容易得出片面的结论。

发明内容

20 鉴于以上内容，本申请提供一种文本情绪分析方法、服务器及计算机可
读存储介质，其主要目的在于提高目标文本情绪分析的准确性及效率。

为实现上述目的，本申请提供一种文本情绪分析方法，该方法包括：

S1、接收携带目标文本的文本情绪分析请求，对所述目标文本进行预处理，
并采用预设序列标注法对经过预处理后的目标文本进行分词处理，得到
25 该目标文本对应的可用词集合；

S2、确定该目标文本的待分析的句子，根据该目标文本对应的可用词集
合获取所述目标文本中待分析的句子对应的可用词集合，根据预设计算规则，
分别计算所述待分析的句子句子向量，其中，句子向量包括句子级别的向
量和词语级别的向量；及

30 S3、分别将所述待分析的句子句子向量输入预先训练好的情绪判断模

型，根据模型输出结果判断所述目标文本的情绪极性。

此外，本申请还提供一种电子装置，其特征在于，该装置包括：存储器、处理器，所述存储器上存储有可在所述处理器上运行的文本情绪分析程序，所述文本情绪分析程序被所述处理器执行时，可实现如下步骤：

5 A1、接收携带目标文本的文本情绪分析请求，对所述目标文本进行预处理，并采用预设序列标注法对经过预处理后的目标文本进行分词处理，得到该目标文本对应的可用词集合；

 A2、确定该目标文本的待分析的句子，根据该目标文本对应的可用词集合获取所述目标文本中待分析的句子对应的可用词集合，根据预设计算规则，
10 分别计算所述待分析的句子句子向量，其中，句子向量包括句子级别的向量和词语级别的向量；及

 A3、分别将所述待分析的句子句子向量输入预先训练好的情绪判断模型，根据模型输出结果判断所述目标文本的情绪极性。

 此外，为实现上述目的，本申请还提供一种计算机可读存储介质，所述
15 计算机可读存储介质中包括文本情绪分析程序，所述文本情绪分析程序被处理器执行时，可实现如上所述文本情绪分析方法中的任意步骤。

 本申请提出的文本情绪分析方法、电子装置及计算机可读存储介质，通过对待分析的目标文本进行分词，根据目标文本的篇幅长短确定待分析的句子，并计算得到目标文本中每个待分析的句子对应的预设类型的句子向量，
20 使得到的句子向量能够更准确的表达待分析的句子信息；利用每个待分析的句子句子向量及预先训练好的情绪判断模型，更准确地判断每个待分析的句子的情绪极性；根据目标文本中每个待分析的句子的情绪极性，综合判断目标文本的情绪极性，有助于提高目标文本情绪分析的准确性；通过从目标文本中筛选出能充分体现目标文本观点的待分析的句子，有助于减小目标
25 文本情绪分析的计算量，提高目标文本情绪分析的效率。

附图说明

图1为本申请文本情绪分析方法较佳实施例的流程图；

图2为本申请电子装置较佳实施例的示意图；

30 图3为本申请图2中文本情绪分析程序的程序模块示意图。

本申请目的的实现、功能特点及优点将结合实施例，参照附图做进一步说明。

具体实施方式

5 应当理解，此处所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。

本申请提供一种文本情绪分析方法。参照图 1 所示，为本申请文本情绪分析方法较佳实施例的流程图。该方法可以由一个装置执行，该装置可以由软件和/或硬件实现。

10 在本实施例中，文本情绪分析方法包括步骤 S1-S3：

S1、接收携带目标文本的文本情绪分析请求，对所述目标文本进行预处理，并采用预设序列标注法对经过预处理后的目标文本进行分词处理，得到该目标文本对应的可用词集合；

15 所述目标文本是关于特定事件/特定人物的文本，文本内容既可能是中文，也可能是英文。例如，当目标文本为国内的各券商、机构发布的研究报告时，鉴于中文不像英语等拉丁语系，除了标点符号之外，字之间紧密相连，没有明显的词边界，因此很难将词简单准确的提取出来。在中文中，单字作为最基本的语义单位，虽然也有自己的意义，但表意能力较差，意义较分散，而词的表意能力更强，能更加准确的描述一个事物，因此在自然语言处理中，
20 通常情况下词（包括单字成词）是最基本的处理单位，因此必须对文本进行准确的分词处理；相反，当目标文本的内容为英文时，例如，X 先生在 twitter 上发布的所有关于中美贸易战的言论，英文的每个单词之间有明显的词边界（空格），则不需做文本分词处理。

25 分词方法大致分为两种：基于词典的机械切分，基于统计模型的序列标注切分两种方式。

在本实施例中，采用序列标注法训练一个长短期记忆循环神经网络(Long Short-Term Memory, 简称 LSTM) 模型作为分词模型进行分词，该分词模型的训练过程如下：

30 获取预设数量（例如，10 万）的样本语句，其中，样本语句中的词语为预设语料库中采用预设序列标注法标注好的词语。其中，该预设语料库使用

的是经典的 bakeoff2005 中的微软研究院的切分语料，可将其中的训练集部分拿过来做训练，将测试集部分作为最终的测试。该预设序列标注法为按照每个字在词中的位置进行标注，标注类型包括：首字标注、中间字标注、尾字标注、单字标注。例如，在一段文字中，可以将每个字按照在词中的位置进行标注，常用的标记有以下四个 label：B，Begin，表示这个字是一个词的首字；M，Middle，表示这是一个词中间的字；E，End，表示这是一个词的尾字；S，Single，表示这是单字成词。分词的过程就是将一段字符输入模型，然后得到相应的标记序列，再根据标记序列进行分词。例如，“达观数据是企业大数据服务商”，经过模型后得到的理想标注序列是：
5 “BMMESBEBMEBME”，最终还原的分词结果是“达观数据/是/企业/大数据/服务商”。

在训练过程中，每隔预设时间，使用训练得到的分词模型对所述测试集中样本语句的各个词语标注进行识别，并将识别字符与该样本语句中采用预设序列标注法的标注进行比对，以评估模型识别标注的误差；若训练得到的模型识别标注的误差出现发散，则调整预设的训练参数并重新训练（例如，
15 利用反向传播算法计算模型误差，并根据误差调整模型参数），直至训练得到的模型识别标注的误差能够收敛；若训练得到的模型识别标注的误差收敛，则结束模型训练，将生成的模型作为分词模型。

需要说明的是，为了保证上述分词步骤顺利进行，在进行分词操作之前，
20 该步骤还包括：将所述目标文本由原始格式转换成目标格式。其中，目标格式为可执行分词操作的格式。例如，当接收到的目标文本为各券商、机构发布的研究报告时，研究报告一般为 pdf 格式，无法直接进行分词操作，因此，将格式为 pdf 的研究报告通过软件转换成可执行分词操作的格式，例如，word。

进一步地，在进行分词操作之前，需要对上述转换格式后的目标文本进行
25 预处理，例如，根据所述目标文本中的句号将所述目标文本划分为多个句子，然后针对每个句子进行分词操作。

S2、确定该目标文本的待分析的句子，根据该目标文本对应的可用词集合获取所述目标文本中待分析的句子对应的可用词集合，根据预设计算规则，分别计算所述待分析的句子句子向量，其中，句子向量包括句子级别的向
30 量和词语级别的向量；

其中，所述“确定该目标文本的待分析的句子”的步骤包括：

统计所述目标文本的字数；当字数小于预设阈值时，获取所述目标文本中的每一个句子作为待分析的句子；当字数大于或等于预设阈值时，则分别获取所述目标文本对应的可用词集合中每个词语的统计特征，根据预设评分规则计算每个词语的评分，筛选出评分排序靠前的词语作为目标文本的关键词，并从所述目标文本中筛选出包含关键词的句子，将所述包含关键词的句子作为待分析的句子。

在本实施例中，通过统计目标文本的字数，确定目标文本篇幅的长短。

当字数小于预设阈值（例如，300）时，目标文本的篇幅较短，目标文本的每个句子都有可能是关键语句，因此，可将目标文本的每个句子作为待分析的句子。当字数大于或等于预设阈值时，目标文本的篇幅较长，目标文本中会存在较多无法表达关键信息的噪声语句，将影响后续情绪分析结果，因此，需从目标文本中筛选出能代表目标文本关键信息的句子进行后续操作。

优选地，根据无监督的统计类方法对每个词语进行打分排序来提取出关键词。具体地，先从目标文本的分词结果中过滤掉重复词语及无意义词语，例如，将目标文本的每个句子的可用词集合中的所有词语提取成一个大的可用词集合（集合中词语不重复，即每个词在这个集合中只出现一次），然后删除一些无意义的词，例如，“我、你、的、是”等。将这些代词、介词等无具体意义的词语预先按经验组成一个列表，然后删除这些无意义的词。

具体地，所述统计特征包括：词频、位置信息及词跨度。

词频，表示一个词在文本中出现的频率。一般情况下，如果一个词在文本中出现的越是频繁，那么这个词就越有可能作为文章的核心词。因此，频率越高，词频评分也越高；

位置信息，一般情况下，词出现的位置对于词来说有着很大的价值。例如，标题、摘要本身就是作者概括出的文章的中心思想，因此出现在这些地方的词具有一定的代表性，更可能成为关键词；

本实施例中，采用的比例是 5：5：1 的方式，为词语的位置信息设定在开头、结尾、中间的重要程度，开头、结尾、中间的划分比例为 1:1:8。例如，如果一个目标文本一共由一万个词语有序组成，位于前百分之十位置的词语划分为开头位置，位于最后百分之十位置的词语划分为结尾位置，其余作为

中间位置，重要程度为 5: 5: 1。例如，假设“试验区”一词一并在目标文本中出现过 5 次，2 次在开头，1 次在结尾，其余在中间，于是‘试验区’一词关于位置的分数为： $5 \times 2 + 5 \times 1 + 1 \times 2 = 17$ 。

词跨度，指一个词或者短语字文中首次出现和末次出现之间的距离，词跨度越大说明这个词对文本越重要，可以反映文本的主题。因此，词跨度越大，词跨度评分也越高。具体地，词跨度的计算公式为：

$$\text{span}_i = (\text{last}_i - \text{first}_i + 1) / \text{sum}$$

其中， last_i 表示词语 i 在目标文本中最后一次出现的位置， first_i 表示词语 i 在目标文本中第一次出现的位置， sum 表示目标文本中词语的总数。词跨度被作为提取关键词的方法是因为在现实中，文本中总是有很多噪声（指不是关键词的那些词），使用词跨度可以减少这些噪声。

综合考虑上述统计特征，并计算可用词集合中每个词语的评分。具体地，所述每个词语的评分的计算公式为：

$$S = \alpha * X_1 + \beta * X_2 + \gamma * X_3$$

其中， X_1 为词语在所述目标文本中出现的频率的词频评分， α 为预设的词频权重， X_2 为词语在所述目标文本中出现位置的位置评分， β 为预设的位置权重， X_3 为词语在所述目标文本中的词跨度评分， γ 为预设的词跨度权重。

根据每个词语的评分高低顺序进行排序，选择排序靠前的 K (K 的范围是个可自选的范围，根据需求来设定) 个词语作为目标文本的关键词。上述步骤综合考虑每个词语的词频、位置信息、词跨度，提高了关键词提取的准确性。然后，从目标文本中筛选出包含上述关键词的句子作为待分析的句子。通过从篇幅较长的目标文本中筛选出包含关键词的句子作为待分析的句子，有助于减小计算量，提高文本情绪分析的效率。

确定目标文本的待分析的句子之后，需根据每个待分析的句子对应的可用词集合计算其对应的句子向量，首先，分别计算每个待分析的句子对应的可用词集合中每个词语的词向量，具体地，该步骤包括：

把每个词语输入预先训练好的词向量模型 (word2vec 模型)，生成一个词语级别 (word-level) 的向量 r^{word} ；把组成每个词语的字母/字符输入预先训练好的卷积神经网络模型 (Convolutional Neural Network, CNN)，生成该词语对应的字母/字符级别 (character-level) 的向量 r^{ch} ；将所述 word-level 的向量

和 character-level 的向量组合得到一个新的词向量 $u_n = [r^{wrd}, r^{wch}]$, 作为每个词语的词向量。

其中, r^{wrd} 表示利用 word2vec 模型训练得到的向量, 其处理方式与现有的 word2vec 模型一致, 这里不再赘述; r^{wch} 表示通过一层卷积神经网络训练得到的向量, 具体步骤如下:

假设词 w 由 M 个字母组成, 每个字母通过一个字符向量矩阵 (character embedding matrix) 转化成一个向量 r^{chr} , 即 $r^{chr} = W^{chr} v^c$, 其中 v^c 是 one-hot 向量 (一个长度为 n 的数组, 只有一个元素是 1.0, 其他元素是 0.0), 依次处理后词 w 就能表示成一个 $d^{chr} * M$ 的向量矩阵了。然后使用一个卷积长度为 k^{chr} 的过滤器对上述向量进行卷积, 再使用一个最大池化层进行池化后得到一个长度为 cl_u^0 的向量, 也即 r^{wch} 。需要说明的是, 本申请方案中卷积的做法和传统卷积不太一样, 把相邻几个向量进行拼接之后, 通过一个线性计算转化成一个固定维度的向量, 统一不同长度词的维度。

通过上述步骤获取一个词语的 word-level 的向量和 character-level 的向量, 有利于同时捕捉到词语的语义信息和词形态信息, 为后面计算每个待分析的句子向量的步骤打下基础。

利用上述步骤获取每个待分析的句子中不同词语的词向量后, 需进一步确定每个句子对应的句子向量。具体地, 该步骤包括: 把每个待分析的句子输入所述 word2vec 模型, 生成一个句子级别 (sentence-level) 的向量;

获取每个待分析的句子对应的可用词集合中每个词语的词向量 $u_1, u_2, \dots, u_n$, 把组成每个待分析的句子词语输入所述卷积神经网络模型, 生成每个待分析的句子对应的 word-level 的向量; 及

将所述 sentence-level 的向量和 word-level 的向量组合得到一个新的句子向量, 作为每个待分析的句子向量的句子向量。

其中, 所述每个待分析的句子对应的 word-level 的向量的计算方法与上述 r^{wch} 的步骤大致相同, 这里不作赘述。

利用上述步骤计算每个待分析的句子向量的句子向量, 使得到的句子向量能更准确地表达每个待分析的句子信息, 为后面判断目标文本的情绪极性打下基础。

S3、分别将所述待分析的句子向量的句子向量输入预先训练好的情绪判断模型, 根据模型输出结果判断所述目标文本的情绪极性。

具体地，预先构建样本库，利用样本库训练预设深度神经网络模型（例如，一个三层的神经网络），确定模型参数，将确定模型参数的神经网络模型作为情绪判断模型。所述情绪判断模型的训练步骤包括：

5 获取预设数量的样本语句，并根据其情绪极性为每个样本语句标注标签，得到样本数据。其中，标签包括：“1”、“0”、“-1”，“1”表示样本语句的情绪极性偏向于正面，“0”表示样本语句的情绪极性偏向于中立，“-1”表示样本语句的情绪极性偏向于负面。

10 基于交叉验证法（cross-validation）将预设数量（例如，10 万）的样本语句按照预设比例（例如，7:1:2）分为：训练集，评估集，测试集这三个部分，其中，测试集为完全不参与模型训练的数据，仅仅是用来观测训练效果的数据；将训练集的样本数据输入至三层的神经网络模型，对该神经网络模型进行训练，初步确定模型参数；为了相对客观地判断初步确定的模型参数对训练集之外的样本数据的拟合程度，将所述测试集的样本数据输入训练得到的所述神经网络模型中，以对训练得到的所述神经网络模型进行测试，当训练
15 得到的所述神经网络模型满足预设验证条件（例如，模型预测准确率大于或等于预设阈值，95%），则训练完成，并将训练完成的神经网络模型作为情绪判断模型。

20 在本实施例中，训练模型时采用的是标注了标签为离散值的样本数据，因此，将目标文本的每个所述待分析的句子输入所述情绪判断模型后，模型输出结果也为离散值。

进一步地，需根据模型输出结果判断所述目标文本的情绪极性。具体地，该步骤包括：

25 根据模型输出结果分别确定每个所述待分析的句子的情绪极性，分别统计不同情绪极性对应的待分析的句子数量；选择待分析的句子数量最多的情绪极性作为目标文本对应的情绪极性。

30 在本实施例中，将每个待分析的句子作为情绪判断模型的输入，输出每个待分析的句子对应的情绪标签，例如，“1”、“0”、“-1”，根据情绪标签确定每个待分析的句子的情绪极性。然后，将所有待分析的句子输出结果进行融合，得到目标文本的情绪极性。在本实施例中，分统计不同情绪极性对应的待分析的句子数量，哪个情绪极性对应的句子数量最多，则

将该情绪极性作为目标文本的情绪极性。例如，通过上述步骤确定 X 先生发布的关于中美贸易战的文本中每一句话的情绪极性、并对不同情绪极性对应的句子进行统计，若“负面”情绪极性对应的句子数量最多，判断 X 先生对中美贸易战的态度为“负面”。

5 上述实施例提出的文本情绪分析方法，通过对待分析的目标文本进行分词，根据目标文本的篇幅长短确定待分析的句子，并计算得到目标文本中每个待分析的句子对应的预设类型的句子向量，使得到的句子向量能够更准确的表达待分析的句子信息；利用每个待分析的句子向量及预先训练好的情绪判断模型，更准确地判断每个待分析的句子的情绪极性；根据目标
10 文本中每个待分析的句子的情绪极性，综合判断目标文本的情绪极性，有助于提高目标文本情绪分析的准确性；通过从目标文本中筛选出能充分体现目标文本观点的待分析的句子，有助于减小目标文本情绪分析的计算量，提高目标文本情绪分析的效率。

15 本申请还提供一种电子装置。参照图 2 所示，为本申请电子装置 1 较佳实施例的示意图。

在本实施例中，电子装置 1 可以是服务器、智能手机、平板电脑、便携计算机、桌上型计算机等具有数据处理功能的终端设备，所述服务器可以是机架式服务器、刀片式服务器、塔式服务器或机柜式服务器。

20 该电子装置 1 包括存储器 11、处理器 12，通信总线 13，及网络接口 14。

其中，存储器 11 至少包括一种类型的可读存储介质，所述可读存储介质包括闪存、硬盘、多媒体卡、卡型存储器（例如，SD 或 DX 存储器等）、磁性存储器、磁盘、光盘等。存储器 11 在一些实施例中可以是所述电子装置 1 的内部存储单元，例如该电子装置 1 的硬盘。存储器 11 在另一些实施例中也可以
25 是所述电子装置 1 的外部存储设备，例如该电子装置 1 上配备的插接式硬盘，智能存储卡（Smart Media Card, SMC），安全数字（Secure Digital, SD）卡，闪存卡（Flash Card）等。进一步地，存储器 11 还可以既包括该电子装置 1 的内部存储单元也包括外部存储设备。

存储器 11 不仅可以用于存储安装于该电子装置 1 的应用软件及各类数据，
30 例如文本情绪分析程序 10 等，还可以用于暂时地存储已经输出或者将要输出

的数据。

处理器 12 在一些实施例中可以是一中央处理器 (Central Processing Unit, CPU)、控制器、微控制器、微处理器或其他数据处理芯片, 用于运行存储器 11 中存储的程序代码或处理数据, 例如文本情绪分析程序 10 等。

5 通信总线 13 用于实现这些组件之间的连接通信。

网络接口 14 可选的可以包括标准的有线接口、无线接口(如 WI-FI 接口), 通常用于在该电子装置 1 与其他电子设备之间建立通信连接。

图 2 仅示出了具有组件 11-14 的电子装置 1, 本领域技术人员可以理解的是, 图 2 示出的结构并不构成对电子装置 1 的限定, 可以包括比图示更少或者更多的部件, 或者组合某些部件, 或者不同的部件布置。

10 可选地, 该电子装置 1 还可以包括用户接口, 用户接口可以包括显示器 (Display)、输入单元比如键盘 (Keyboard), 可选的用户接口还可以包括标准的有线接口、无线接口。

15 可选地, 在一些实施例中, 显示器可以是 LED 显示器、液晶显示器、触控式液晶显示器以及有机发光二极管 (Organic Light-Emitting Diode, OLED) 触摸器等。其中, 显示器也可以称为显示屏或显示单元, 用于显示在电子装置 1 中处理的信息以及用于显示可视化的用户界面。

在图 2 所示的电子装置 1 实施例中, 作为一种计算机存储介质的存储器 20 11 中存储文本情绪分析程序 10 的程序代码, 处理器 12 执行文本情绪分析程序 10 的程序代码时, 实现如下步骤:

A1、接收携带目标文本的文本情绪分析请求, 对所述目标文本进行预处理, 并采用预设序列标注法对经过预处理后的目标文本进行分词处理, 得到该目标文本对应的可用词集合;

25 所述目标文本是关于特定事件/特定人物的文本, 文本内容既可能是中文, 也可能是英文。例如, 当目标文本为国内的各券商、机构发布的研究报告时, 需对文本进行准确的分词处理; 相反, 当目标文本的内容为英文时, 例如, X 先生在 twitter 上发布的所有关于中美贸易战的言论, 英文的每个单词之间有明显的词边界 (空格), 则不需做文本分词处理。

30 分词方法大致分为两种: 基于词典的机械切分, 基于统计模型的序列标

注切分两种方式。

在本实施例中，采用序列标注法训练一个长短期记忆循环神经网络（LSTM）模型作为分词模型进行分词，该分词模型的训练过程如下：

获取预设数量（例如，10 万）的样本语句，其中，样本语句中的词语为
5 预设语料库中采用预设序列标注法标注好的词语。其中，该预设语料库使用的是经典的 bakeoff2005 中的微软研究院的切分语料，可将其中的训练集部分拿过来做训练，将测试集部分作为最终的测试。该预设序列标注法为按照每个字在词中的位置进行标注，标注类型包括：首字标注、中间字标注、尾字标注、单字标注。例如，在一段文字中，可以将每个字按照在词中的位置进
10 行标注，常用的标记有以下四个 label：B，Begin，表示这个字是一个词的首字；M，Middle，表示这是一个词中间的字；E，End，表示这是一个词的尾字；S，Single，表示这是单字成词。分词的过程就是将一段字符输入模型，然后得到相应的标记序列，再根据标记序列进行分词。例如，“达观数据是企业大数据服务商”，经过模型后得到的理想标注序列是：
15 “BMMESBEBMEBME”，最终还原的分词结果是“达观数据/是/企业/大数据/服务商”。

在训练过程中，每隔预设时间，使用训练得到的分词模型对所述测试集中样本语句的各个词语标注进行识别，并将识别字符与该样本语句中采用预设序列标注法的标注进行比对，以评估模型识别标注的误差；若训练得到的
20 模型识别标注的误差出现发散，则调整预设的训练参数并重新训练（例如，利用反向传播算法计算模型误差，并根据误差调整模型参数），直至训练得到的模型识别标注的误差能够收敛；若训练得到的模型识别标注的误差收敛，则结束模型训练，将生成的模型作为分词模型。

需要说明的是，为了保证上述分词步骤顺利进行，在进行分词操作之前，
25 该步骤还包括：将所述目标文本由原始格式转换成目标格式。其中，目标格式为可执行分词操作的格式。例如，当接收到的目标文本为各券商、机构发布的研究报告时，研究报告一般为 pdf 格式，无法直接进行分词操作，因此，将格式为 pdf 的研究报告通过软件转换成可执行分词操作的格式，例如，word。

进一步地，在进行分词操作之前，需要对上述转换格式后的目标文本进
30 行预处理，例如，根据所述目标文本中的句号将所述目标文本划分为多个句

子，然后针对每个句子进行分词操作。

A2、确定该目标文本的待分析的句子，根据该目标文本对应的可用词集合获取所述目标文本中待分析的句子对应的可用词集合，根据预设计算规则，分别计算所述待分析的句子句子向量，其中，句子向量包括句子级别的向量和词语级别的向量；

其中，所述“确定该目标文本的待分析的句子”的步骤包括：

统计所述目标文本的字数；当字数小于预设阈值时，获取所述目标文本中的每一个句子作为待分析的句子；当字数大于或等于预设阈值时，则分别获取所述目标文本对应的可用词集合中每个词语的统计特征，根据预设评分规则计算每个词语的评分，筛选出评分排序靠前的词语作为目标文本的关键词，并从所述目标文本中筛选出包含关键词的句子，将所述包含关键词的句子作为待分析的句子。

在本实施例中，通过统计目标文本的字数，确定目标文本篇幅的长短。

当字数小于预设阈值（例如，300）时，目标文本的篇幅较短，目标文本的每个句子都有可能是关键语句，因此，可将目标文本的每个句子作为待分析的句子。当字数大于或等于预设阈值时，目标文本的篇幅较长，目标文本中会存在较多无法表达关键信息的噪声语句，将影响后续情绪分析结果，因此，需从目标文本中筛选出能代表目标文本关键信息的句子进行后续操作。

优选地，根据无监督的统计类方法对每个词语进行打分排序来提取出关键词。具体地，先从目标文本的分词结果中过滤掉重复词语及无意义词语，例如，将目标文本的每个句子的可用词集合中的所有词语提取成一个大的可用词集合（集合中词语不重复，即每个词在这个集合中只出现一次），然后删除一些无意义的词，例如，“我、你、的、是”等。将这些代词、介词等无具体意义的词语预先按经验组成一个列表，然后删除这些无意义的词。

具体地，所述统计特征包括：词频、位置信息及词跨度。

词频，表示一个词在文本中出现的频率，频率越高，词频评分也越高。

位置信息，一般情况下，词出现的位置对于词来说有着很大的价值，本实施例中，采用的比例是 5: 5: 1 的方式，为词语的位置信息设定在开头、结尾、中间的重要程度，开头、结尾、中间的划分比例为 1:1:8。

词跨度，指一个词或者短语字文中首次出现和末次出现之间的距离，词

跨度越大说明这个词对文本越重要，可以反映文本的主题。因此，词跨度越大，词跨度评分也越高。具体地，词跨度的计算公式为：

$$\text{span}_i = (\text{last}_i - \text{first}_i + 1) / \text{sum}$$

其中， last_i 表示词语 i 在目标文本中最后一次出现的位置， first_i 表示词语 i 在目标文本中第一次出现的位置， sum 表示目标文本中词语的总数。词跨度
5 被作为提取关键词的方法是因为在现实中，文本中总是有很多噪声（指不是关键词的那些词），使用词跨度可以减少这些噪声。

综合考虑上述统计特征，并计算可用词集合中每个词语的评分。具体地，所述每个词语的评分的计算公式为：

$$S = \alpha * X_1 + \beta * X_2 + \gamma * X_3$$

其中， X_1 为词语在所述目标文本中出现的频率的词频评分， α 为预设的词
10 频权重， X_2 为词语在所述目标文本中出现位置的位置评分， β 为预设的位置权重， X_3 为词语在所述目标文本中的词跨度评分， γ 为预设的词跨度权重。

根据每个词语的评分高低顺序进行排序，选择排序靠前的 K （ K 的范围是个可自选的范围，根据需求来设定）个词语作为目标文本的关键词。上述
15 步骤综合考虑每个词语的词频、位置信息、词跨度，提高了关键词提取的准确性。然后，从目标文本中筛选出包含上述关键词的句子作为待分析的句子。通过从篇幅较长的目标文本中筛选出包含关键词的句子作为待分析的句子，有助于减小计算量，提高文本情绪分析的效率。

确定目标文本的待分析的句子之后，需根据每个待分析的句子对应的可用词集合计算其对应的句子向量，首先，分别计算每个待分析的句子对应的
20 可用词集合中每个词语的词向量，具体地，该步骤包括：

把每个词语输入预先训练好的词向量模型（word2vec 模型），生成一个词语级别（word-level）的向量 r^{word} ；把组成每个词语的字母/字符输入预先训练好的卷积神经网络模型（Convolutional Neural Network, CNN），生成该词语对应的字母/字符级别（character-level）的向量 r^{wch} ；将所述词语级别（word-level）
25 的向量和字母/字符级别（character-level）的向量组合得到一个新的词向量 $u_n = [r^{\text{word}}, r^{\text{wch}}]$ ，作为每个词语的词向量。

其中， r^{word} 表示利用 word2vec 模型训练得到的向量，其处理方式与现有的 word2vec 模型一致，这里不再赘述； r^{wch} 表示通过一层卷积神经网络训练

得到的向量，具体步骤如下：

假设词 w 由 M 个字母组成，每个字母通过一个字符向量矩阵（character embedding matrix）转化成一个向量 r^{chr} ，即 $r^{chr} = W^{chr} v^c$ ，其中 v^c 是 one-hot 向量（一个长度为 n 的数组，只有一个元素是 1.0，其他元素是 0.0），依次处
5 理后词 w 就能表示成一个 $d^{chr} * M$ 的向量矩阵了。然后使用一个卷积长度为 k^{chr} 的过滤器对上述向量进行卷积，再使用一个最大池化层进行池化后得到一个长度为 cl_u^0 的向量，也即 r^{wch} 。需要说明的是，本申请方案中卷积的做法和传统卷积不太一样，把相邻几个向量进行拼接之后，通过一个线性计算转化成一个固定维度的向量，统一不同长度词的维度。

10 通过上述步骤获取一个词语的 word-level 的向量和 character-level 的向量，有利于同时捕捉到词语的语义信息和词形态信息，为后面计算每个待分析的句子句子向量的步骤打下基础。

利用上述步骤获取每个待分析的句子中不同词语的词向量后，需进一步确定每个句子对应的句子向量。具体地，该步骤包括：

15 把每个待分析的句子输入所述 word2vec 模型，生成一个句子级别（sentence-level）的向量；获取每个待分析的句子对应的可用词集合中每个词语的词向量 $u_1, u_2, \dots, u_n$ ，把组成每个待分析的句子词语输入所述卷积神经网络模型，生成每个待分析的句子对应的 word-level 的向量；将所述 sentence-level 的向量和 word-level 的向量组合得到一个新的句子向量，作为每
20 个待分析的句子句子向量。

其中，所述每个待分析的句子对应的 word-level 的向量的计算方法与上述 r^{wch} 的步骤大致相同，这里不作赘述。

利用上述步骤计算每个待分析的句子句子向量，使得到的句子向量能更准确地表达每个待分析的句子信息，为后面判断目标文本的情绪极性打
25 下基础。

A3、分别将所述待分析的句子句子向量输入预先训练好的情绪判断模型，根据模型输出结果判断所述目标文本的情绪极性。

具体地，预先构建样本库，利用样本库训练预设深度神经网络模型（例如，一个三层的神经网络），确定模型参数，将确定模型参数的神经网络模型
30 作为情绪判断模型。所述情绪判断模型的训练步骤包括：

获取预设数量的样本语句，并根据其情绪极性为每个样本语句标注标签，

得到样本数据。其中，标签包括：“1”、“0”、“-1”，“1”表示样本语句的情绪极性偏向于正面，“0”表示样本语句的情绪极性偏向于中立，“-1”表示样本语句的情绪极性偏向于负面。

基于交叉验证法（cross-validation）将预设数量（例如，10 万）的样本语句按照预设比例（例如，7:1:2）分为：训练集、评估集、测试集这三个部分，其中，测试集为完全不参与模型训练的数据，仅仅是用来观测训练效果的数据；将训练集的样本数据输入至三层的神经网络模型，对该神经网络模型进行训练，初步确定模型参数；为了相对客观地判断初步确定的模型参数对训练集之外的样本数据的拟合程度，将所述测试集的样本数据输入训练得到的所述神经网络模型中，以对训练得到的所述神经网络模型进行测试，当训练得到的所述神经网络模型满足预设验证条件（例如，模型预测准确率大于或等于预设阈值，95%），则训练完成，并将训练完成的神经网络模型作为情绪判断模型。

在本实施例中，训练模型时采用的是标注了标签为离散值的样本数据，因此，将目标文本的每个所述待分析的句子输入所述情绪判断模型后，模型输出结果也为离散值。

进一步地，需根据模型输出结果判断所述目标文本的情绪极性。具体地，该步骤包括：

根据模型输出结果分别确定每个所述待分析的句子的情绪极性，分别统计不同情绪极性对应的待分析的句子数量；选择待分析的句子数量最多的情绪极性作为目标文本对应的情绪极性。

将每个待分析的句子作为情绪判断模型的输入，输出每个待分析的句子对应的情绪标签，例如，“1”、“0”、“-1”，根据情绪标签确定每个待分析的句子的情绪极性。然后，将所有待分析的句子输出结果进行融合，得到目标文本的情绪极性。在本实施例中，分别统计不同情绪极性对应的待分析的句子数量，哪个情绪极性对应的句子数量最多，则将该情绪极性作为目标文本的情绪极性。例如，通过上述步骤确定 X 先生发布的关于中美贸易战的文本中每一句话的情绪极性、并对不同情绪极性对应的句子进行统计，若“负面”情绪极性对应的句子数量最多，判断 X 先生对中美贸易战的态度为“负面”。

上述实施例提出的电子装置 1，通过对待分析的目标文本进行分词，根据目标文本的篇幅长短确定待分析的句子，并计算得到目标文本中每个待分析的句子对应的预设类型的句子向量，使得到的句子向量能够更准确的表达待分析的句子信息；利用每个待分析的句子向量及预先训练好的情绪判断模型，更准确地判断每个待分析的句子的情绪极性；根据目标文本中每个待分析的句子的情绪极性，综合判断目标文本的情绪极性，有助于提高目标文本情绪分析的准确性；通过从目标文本中筛选出能充分体现目标文本观点的待分析的句子，有助于减小目标文本情绪分析的计算量，提高目标文本情绪分析的效率。

可选地，在其他的实施例中，文本情绪分析程序 10 还可以被分割为一个或者多个模块，一个或者多个模块被存储于存储器 11 中，并由一个或多个处理器（本实施例为处理器 12）所执行，以完成本申请，本申请所称的模块是指能够完成特定功能的一系列计算机程序指令段。例如，参照图 3 所示，为图 2 中文本情绪分析程序 10 的模块示意图，该实施例中，文本情绪分析程序 10 可以被分割为分词模块 110、向量计算模块 120 及情绪分析模块 130，所述模块 110-130 所实现的功能或操作步骤均与上文类似，此处不再详述，示例性地，例如其中：

分词模块 110，用于接收携带目标文本的文本情绪分析请求，对所述目标文本进行预处理，并采用预设序列标注法对经过预处理后的目标文本进行分词处理，得到该目标文本对应的可用词集合；

向量计算模块 120，用于确定该目标文本的待分析的句子，根据该目标文本对应的可用词集合获取所述目标文本中待分析的句子对应的可用词集合，根据预设计算规则，分别计算所述待分析的句子向量，其中，句子向量包括句子级别的向量和词语级别的向量；及

情绪分析模块 130，用于分别将所述待分析的句子向量输入预先训练好的情绪判断模型，根据模型输出结果判断所述目标文本的情绪极性。

此外，本申请实施例还提出一种计算机可读存储介质，所述计算机可读存储介质中包括文本情绪分析程序 10，所述文本情绪分析程序 10 被处理器执行时实现如下操作：

A1、接收携带目标文本的文本情绪分析请求，对所述目标文本进行预处理，并采用预设序列标注法对经过预处理后的目标文本进行分词处理，得到该目标文本对应的可用词集合；

5 A2、确定该目标文本的待分析的句子，根据该目标文本对应的可用词集合获取所述目标文本中待分析的句子对应的可用词集合，根据预设计算规则，分别计算所述待分析的句子句子向量，其中，句子向量包括句子级别的向量和词语级别的向量；及

A3、分别将所述待分析的句子句子向量输入预先训练好的情绪判断模型，根据模型输出结果判断所述目标文本的情绪极性。

10 本申请之计算机可读存储介质的具体实施方式与上述文本情绪分析方法的具体实施方式大致相同，在此不再赘述。

上述本申请实施例序号仅仅为了描述，不代表实施例的优劣。

需要说明的是，在本文中，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、装置、物品或者方法15 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、装置、物品或者方法所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括该要素的过程、装置、物品或者方法中还存在另外的相同要素。

通过以上的实施方式的描述，本领域的技术人员可以清楚地了解到上述20 实施例方法可借助软件加必需的通用硬件平台的方式来实现，当然也可以通过硬件，但很多情况下前者是更佳的实施方式。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如上所述的一个存储介质（如 ROM/RAM、磁碟、光盘）中，包括若干指令用以使得一台终端设备（可以是手机，计算机，服务器，或者网络设备等）执行本申请各个实施例所述的方法。25

以上仅为本申请的优选实施例，并非因此限制本申请的专利范围，凡是利用本申请说明书及附图内容所作的等效结构或等效流程变换，或直接或间接运用在其它相关的技术领域，均同理包括在本申请的专利保护范围内。

权 利 要 求 书

1、一种文本情绪分析方法，应用于电子装置，其特征在于，所述方法包括：

5 S1、接收携带目标文本的文本情绪分析请求，对所述目标文本进行预处理，并采用预设序列标注法对经过预处理后的目标文本进行分词处理，得到该目标文本对应的可用词集合；

S2、确定该目标文本的待分析的句子，根据该目标文本对应的可用词集合获取所述目标文本中待分析的句子对应的可用词集合，根据预设计算规则，分别计算所述待分析的句子句子向量，其中，句子向量包括句子级别的向量和词语级别的向量；及

10 S3、分别将所述待分析的句子句子向量输入预先训练好的情绪判断模型，根据模型输出结果判断所述目标文本的情绪极性。

2、根据权利要求 1 所述的文本情绪分析方法，其特征在于，所述“根据模型输出结果判断所述目标文本的情绪极性”的步骤包括：

15 根据模型输出结果分别确定每个所述待分析的句子的情绪极性，分别统计不同情绪极性对应的待分析的句子数量；及

选择待分析的句子数量最多的情绪极性作为目标文本对应的情绪极性。

3、根据权利要求 2 所述的文本情绪分析方法，其特征在于，所述“分别计算所述待分析的句子句子向量”的步骤包括：

20 把所述待分析的句子输入所述词向量模型，生成一个句子级别的向量；

获取所述待分析的句子对应的可用词集合中每个词语的词向量，把组成每个所述待分析的句子词语输入所述卷积神经网络模型，生成每个句子对应的词语级别的向量；及

25 将所述句子级别的向量和词语级别的向量组合得到一个新的句子向量，作为每个所述待分析的句子句子向量。

4、根据权利要求 3 所述的文本情绪分析方法，其特征在于，所述“获取所述待分析的句子对应的可用词集合中每个词语的词向量”的步骤包括：

把每个词语输入预先训练好的词向量模型，生成一个词语级别的向量；

30 把组成每个词语的字母/字符输入预先训练好的卷积神经网络模型，生成该词语对应的字母/字符级别的向量；及

将所述词语级别的向量和字母/字符级别的向量组合得到一个新的词向量，作为每个词语的词向量。

5、根据权利要求 4 所述的文本情绪分析方法，其特征在于，所述“采用预设序列标注法对经过预处理后的目标文本进行分词处理”的步骤包括：

5 对所述经过预处理后的目标文本中的每个字在词语中的位置进行标注，标注类型包括首字标注、中间字标注、尾字标注、单字标注；及

根据每个字的标注类型，确定所述经过预处理后的目标文本的分词结果。

6、根据权利要求 1 至 5 中任意一项所述的文本情绪分析方法，其特征在于，所述“确定该目标文本的待分析的句子”的步骤包括：

10 统计所述目标文本的字数；

当字数小于预设阈值时，获取所述目标文本中的每一个句子作为待分析的句子；或

当字数大于或等于预设阈值时，则分别获取所述目标文本对应的可用词集合中每个词语的统计特征，根据预设评分规则计算每个词语的评分，筛选
15 出评分排序靠前的词语作为目标文本的关键词，并从所述目标文本中筛选出包含关键词的句子，将所述包含关键词的句子作为待分析的句子。

7、根据权利要求 6 所述的文本情绪分析方法，其特征在于：

所述统计特征包括词频、位置信息及词跨度；

所述每个词语的评分的计算公式为：

$$S = \alpha * X_1 + \beta * X_2 + \gamma * X_3$$

20 其中， X_1 为词语在所述目标文本中出现的频率的词频评分， α 为预设的词频权重， X_2 为词语在所述目标文本中出现位置的位置评分， β 为预设的位置权重， X_3 为词语在所述目标文本中的词跨度评分， γ 为预设的词跨度权重。

8、一种电子装置，其特征在于，该装置包括：存储器、处理器，所述存储器上存储有可在所述处理器上运行的文本情绪分析程序，所述文本情绪分
25 析程序被所述处理器执行时，可实现如下步骤：

A1、接收携带目标文本的文本情绪分析请求，对所述目标文本进行预处理，并采用预设序列标注法对经过预处理后的目标文本进行分词处理，得到该目标文本对应的可用词集合；

A2、确定该目标文本的待分析的句子，根据该目标文本对应的可用词集

合获取所述目标文本中待分析的句子对应的可用词集合，根据预设计算规则，分别计算所述待分析的句子句子向量，其中，句子向量包括句子级别的向量和词语级别的向量；及

5 A3、分别将所述待分析的句子句子向量输入预先训练好的情绪判断模型，根据模型输出结果判断所述目标文本的情绪极性。

9、根据权利要求8所述的电子装置，其特征在于，所述“根据模型输出结果判断所述目标文本的情绪极性”的步骤包括：

根据模型输出结果分别确定每个所述待分析的句子的情绪极性，分别统计不同情绪极性对应的待分析的句子数量；及

10 选择待分析的句子数量最多的情绪极性作为目标文本对应的情绪极性。

10、根据权利要求9所述的电子装置，其特征在于，所述“分别计算所述待分析的句子句子向量”的步骤包括：

把所述待分析的句子输入所述词向量模型，生成一个句子级别的向量；

15 获取所述待分析的句子对应的可用词集合中每个词语的词向量，把组成每个所述待分析的句子词语输入所述卷积神经网络模型，生成每个句子对应的词语级别的向量；及

将所述句子级别的向量和词语级别的向量组合得到一个新的句子向量，作为每个所述待分析的句子句子向量。

20 11、根据权利要求10所述的电子装置，其特征在于，所述“获取所述待分析的句子对应的可用词集合中每个词语的词向量”的步骤包括：

把每个词语输入预先训练好的词向量模型，生成一个词语级别的向量；

把组成每个词语的字母/字符输入预先训练好的卷积神经网络模型，生成该词语对应的字母/字符级别的向量；及

25 将所述词语级别的向量和字母/字符级别的向量组合得到一个新的词向量，作为每个词语的词向量。

12、根据权利要求11所述的电子装置，其特征在于，所述“采用预设序列标注法对经过预处理后的目标文本进行分词处理”的步骤包括：

对所述经过预处理后的目标文本中的每个字在词语中的位置进行标注，标注类型包括首字标注、中间字标注、尾字标注、单字标注；及

30 根据每个字的标注类型，确定所述经过预处理后的目标文本的分词结果。

13、根据权利要求 8 至 12 中任意一项所述的电子装置，其特征在于，所述“确定该目标文本的待分析的句子”的步骤包括：

统计所述目标文本的字数；

当字数小于预设阈值时，获取所述目标文本中的每一个句子作为待分析的句子；或

当字数大于或等于预设阈值时，则分别获取所述目标文本对应的可用词集合中每个词语的统计特征，根据预设评分规则计算每个词语的评分，筛选出评分排序靠前的词语作为目标文本的关键词，并从所述目标文本中筛选出包含关键词的句子，将所述包含关键词的句子作为待分析的句子。

14、根据权利要求 8 所述的电子装置，其特征在于：

所述统计特征包括词频、位置信息及词跨度；

所述每个词语的评分的计算公式为：

$$S = \alpha * X_1 + \beta * X_2 + \gamma * X_3$$

其中， X_1 为词语在所述目标文本中出现的频率的词频评分， α 为预设的词频权重， X_2 为词语在所述目标文本中出现位置的位置评分， β 为预设的位置权重， X_3 为词语在所述目标文本中的词跨度评分， γ 为预设的词跨度权重。

15、一种计算机可读存储介质，其特征在于，所述计算机可读存储介质中包括文本情绪分析程序，所述文本情绪分析程序被处理器执行时，可实现如下步骤：

S1、接收携带目标文本的文本情绪分析请求，对所述目标文本进行预处理，并采用预设序列标注法对经过预处理后的目标文本进行分词处理，得到该目标文本对应的可用词集合；

S2、确定该目标文本的待分析的句子，根据该目标文本对应的可用词集合获取所述目标文本中待分析的句子对应的可用词集合，根据预设计算规则，分别计算所述待分析的句子句子向量，其中，句子向量包括句子级别的向量和词语级别的向量；及

S3、分别将所述待分析的句子句子向量输入预先训练好的情绪判断模型，根据模型输出结果判断所述目标文本的情绪极性。

16、根据权利要求 1 所述的计算机可读存储介质，其特征在于，所述“根据模型输出结果判断所述目标文本的情绪极性”的步骤包括：

根据模型输出结果分别确定每个所述待分析的句子的情绪极性，分别统计不同情绪极性对应的待分析的句子数量；及

选择待分析的句子数量最多的情绪极性作为目标文本对应的情绪极性。

17、根据权利要求 16 所述的计算机可读存储介质，其特征在于，所述“分别计算所述待分析的句子”的步骤包括：

把所述待分析的句子输入所述词向量模型，生成一个句子级别的向量；

获取所述待分析的句子对应的可用词集合中每个词语的词向量，把组成每个所述待分析的句子词语输入所述卷积神经网络模型，生成每个句子对应的词语级别的向量；及

10 将所述句子级别的向量和词语级别的向量组合得到一个新的句子向量，作为每个所述待分析的句子。

18、根据权利要求 17 所述的计算机可读存储介质，其特征在于，所述“获取所述待分析的句子”的步骤包括：

把每个词语输入预先训练好的词向量模型，生成一个词语级别的向量；

15 把组成每个词语的字母/字符输入预先训练好的卷积神经网络模型，生成该词语对应的字母/字符级别的向量；及

将所述词语级别的向量和字母/字符级别的向量组合得到一个新的词向量，作为每个词语的词向量。

19、根据权利要求 18 所述的计算机可读存储介质，其特征在于，所述“采用预设序列标注法对经过预处理后的目标文本进行分词处理”的步骤包括：

对所述经过预处理后的目标文本中的每个字在词语中的位置进行标注，标注类型包括首字标注、中间字标注、尾字标注、单字标注；及

根据每个字的标注类型，确定所述经过预处理后的目标文本的分词结果。

20、根据权利要求 15 至 19 中任意一项所述的计算机可读存储介质，其特征在于，所述“确定该目标文本的待分析的句子”的步骤包括：

统计所述目标文本的字数；

当字数小于预设阈值时，获取所述目标文本中的每一个句子作为待分析的句子；或

25 当字数大于或等于预设阈值时，则分别获取所述目标文本对应的可用词集合中每个词语的统计特征，根据预设评分规则计算每个词语的评分，筛选

出评分排序靠前的词语作为目标文本的关键词，并从所述目标文本中筛选出包含关键词的句子，将所述包含关键词的句子作为待分析的句子。

21、根据权利要求 15 所述的计算机可读存储介质，其特征在于：

所述统计特征包括词频、位置信息及词跨度；

5 所述每个词语的评分的计算公式为：

$$S = \alpha * X_1 + \beta * X_2 + \gamma * X_3$$

其中， X_1 为词语在所述目标文本中出现的频率的词频评分， α 为预设的词频权重， X_2 为词语在所述目标文本中出现位置的位置评分， β 为预设的位置权重， X_3 为词语在所述目标文本中的词跨度评分， γ 为预设的词跨度权重。

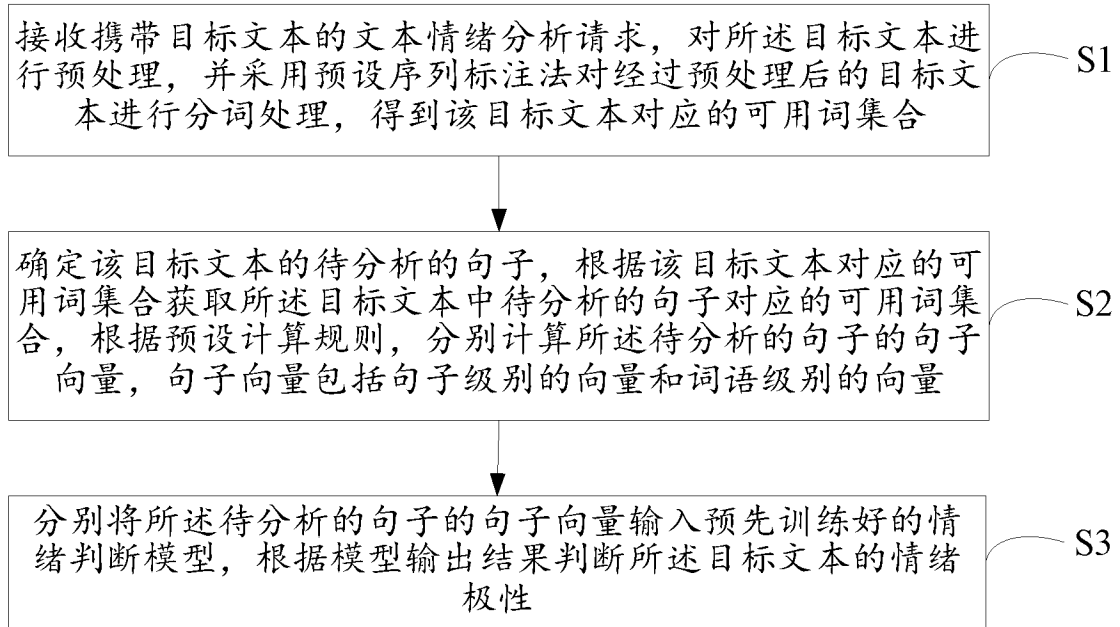


图 1

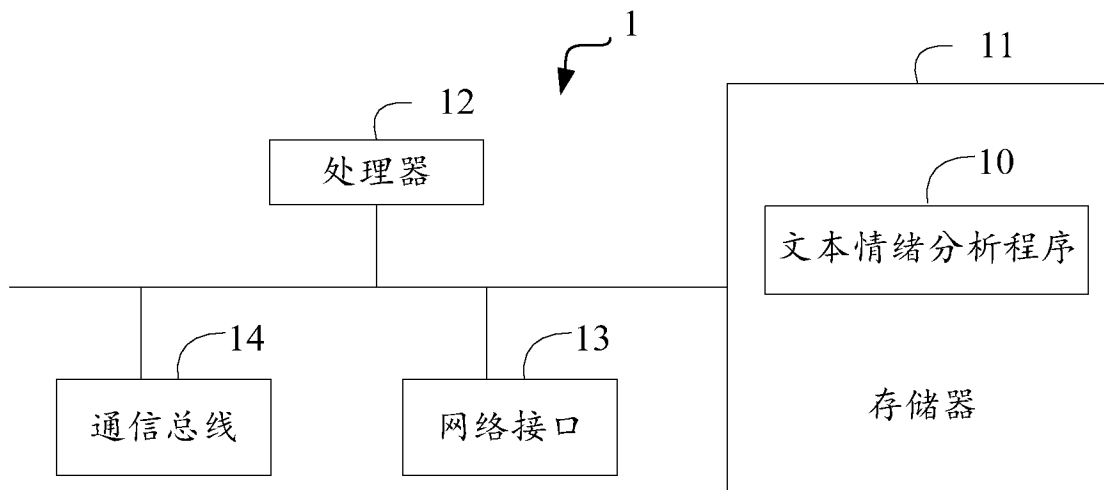


图 2



图 3

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2018/107725

A. CLASSIFICATION OF SUBJECT MATTER

G06F 17/27(2006.01)i; G06N 3/04(2006.01)i; G06K 9/62(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F; G06N; G06K

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

WPI, EPODOC, CNPAT, CNKI, IEEE: 文本, 分词, 序列, 标注, 句子, 向量, 训练, 模型, 情绪, 情感, 极性, text, word, segmentation, sequence, labeling, sentence, vector, train, model, emotional, polarity

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 107944014 A (HOHAI UNIVERSITY) 20 April 2018 (2018-04-20) description, paragraphs [0048]-[0073] and [0087]-[0099]	1-21
A	CN 106919673 A (ZHEJIANG GONGSHANG UNIVERSITY) 04 July 2017 (2017-07-04) entire document	1-21
A	CN 107403017 A (SHANGHAI SHUDAN INFORMATION TECHNOLOGY CO., LTD.) 28 November 2017 (2017-11-28) entire document	1-21
A	CN 106294845 A (TSINGHUA UNIVERSITY) 04 January 2017 (2017-01-04) entire document	1-21
A	CN 107239439 A (TONGJI UNIVERSITY) 10 October 2017 (2017-10-10) entire document	1-21

☐ Further documents are listed in the continuation of Box C.☒ See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“E” earlier application or patent but published on or after the international filing date

“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

“O” document referring to an oral disclosure, use, exhibition or other means

“P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

“&” document member of the same patent family

Date of the actual completion of the international search

11 January 2019

Date of mailing of the international search report

30 January 2019

Name and mailing address of the ISA/CN

State Intellectual Property Office of the P. R. China
No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing
100088
China

Facsimile No. (86-10)62019451

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2018/107725

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN	107944014	A	20 April 2018	None	
CN	106919673	A	04 July 2017	None	
CN	107403017	A	28 November 2017	None	
CN	106294845	A	04 January 2017	None	
CN	107239439	A	10 October 2017	None	

国际检索报告

国际申请号

PCT/CN2018/107725

A. 主题的分类

G06F 17/27(2006.01)i; G06N 3/04(2006.01)i; G06K 9/62(2006.01)i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

G06F; G06N; G06K

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称, 和使用的检索词(如使用))

WPI, EPODOC, CNPAT, CNKI, IEEE: 文本, 分词, 序列, 标注, 句子, 向量, 训练, 模型, 情绪, 情感, 极性, text, word, segmentation, sequence, labeling, sentence, vector, train, model, emotional, polarity

C. 相关文件

类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求
X	CN 107944014 A (河海大学) 2018年 4月 20日 (2018 - 04 - 20) 说明书第[0048]-[0073]、[0087]-[0099]段	1-21
A	CN 106919673 A (浙江工商大学) 2017年 7月 4日 (2017 - 07 - 04) 全文	1-21
A	CN 107403017 A (上海数旦信息技术有限公司) 2017年 11月 28日 (2017 - 11 - 28) 全文	1-21
A	CN 106294845 A (清华大学) 2017年 1月 4日 (2017 - 01 - 04) 全文	1-21
A	CN 107239439 A (同济大学) 2017年 10月 10日 (2017 - 10 - 10) 全文	1-21

☐ 其余文件在C栏的续页中列出。☒ 见同族专利附件。

* 引用文件的具体类型:

“A” 认为不特别相关的表示了现有技术一般状态的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件 (如具体说明的)

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期

2019年 1月 11日

国际检索报告邮寄日期

2019年 1月 30日

ISA/CN的名称和邮寄地址

中国国家知识产权局(ISA/CN)
中国北京市海淀区蓟门桥西土城路6号 100088

传真号 (86-10)62019451

受权官员

王晓敏

电话号码 86-(10)-53961359

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2018/107725

检索报告引用的专利文件			公布日 (年/月/日)	同族专利	公布日 (年/月/日)
CN	107944014	A	2018年 4月 20日	无	
CN	106919673	A	2017年 7月 4日	无	
CN	107403017	A	2017年 11月 28日	无	
CN	106294845	A	2017年 1月 4日	无	
CN	107239439	A	2017年 10月 10日	无	