(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 108804512 B
(45)授权公告日 2020.11.24

(21)申请号 201810361702.0
(22)申请日 2018.04.20
(65)同一申请的已公布的文献号
申请公布号 CN 108804512 A
(43)申请公布日 2018.11.13
(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八卦岭工业区平安大厦六楼
(72)发明人 王健宗 吴天博 黄章成 肖京
(74)专利代理机构 深圳市沃德知识产权代理事务所(普通合伙) 44347
代理人 高杰 于志光
(51)Int.Cl.
G06F 16/35(2019.01)
G06F 16/36(2019.01)

(56)对比文件
CN 105740349 A,2016.07.06
CN 107491531 A,2017.12.19
CN 107122382 A,2017.09.01
CN 106598940 A,2017.04.26
WO 2017202125 A1,2017.11.30

审查员 倪浩天

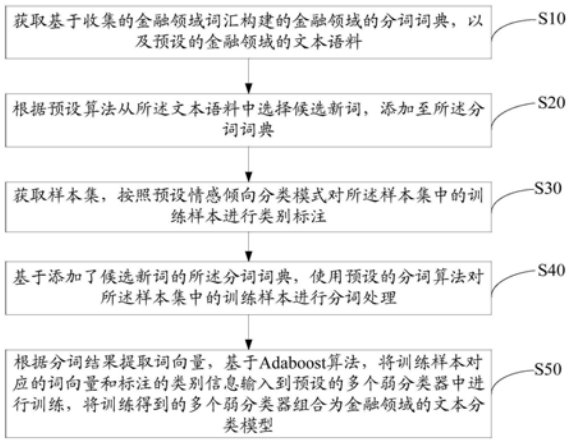
权利要求书3页 说明书10页 附图2页

(54)发明名称

文本分类模型的生成装置、方法及计算机可读存储介质

(57)摘要

本发明公开了一种文本分类模型的生成装置，包括存储器和处理器，存储器上存储有可在处理器上运行的模型生成程序，该程序被处理器执行时实现如下步骤：获取金融领域的分词词典以及金融领域的文本语料；从文本语料中选择候选新词添加至分词词典；获取样本集并对样本集中的训练样本进行类别标注；基于添加了候选新词的分词词典，使用预设的分词算法对样本集中的训练样本进行分词并提取词向量，基于adaboost算法，将词向量和标注的类别信息输入到多个弱分类器中训练，得到文本分类模型。本发明还提出一种文本分类模型的生成方法以及一种计算机可读存储介质。本发明解决了现有技术中无法实现对金融领域文本进行情感倾向性的分类的问题。



CN 108804512 B

1.一种文本分类模型的生成装置，其特征在于，所述装置包括存储器和处理器，所述存储器上存储有可在所述处理器上运行的模型生成程序，所述模型生成程序被所述处理器执行时实现如下步骤：

获取基于收集的金融领域词汇构建的金融领域的分词词典，以及预设的金融领域的文本语料；

根据预设算法从所述文本语料中选择候选新词，添加至所述分词词典，所述根据预设算法从所述文本语料中选择候选新词，添加至所述分词词典的步骤包括：

基于所述分词词典，使用所述分词算法对所述文本语料进行分词处理，根据所述分词结果获取候选词集合；

计算所述候选词集合中各个候选词的信息增益，选择信息增益大于第一预设阈值的候选词作为第一候选新词，将所述第一候选新词添加到所述分词词典中；

基于添加了所述第一候选新词的分词词典，使用所述分词算法对所述文本语料进行分词，并使用分词处理后的文本语料训练词向量模型；

使用训练得到的词向量模型计算分词结果中的词与所述第一候选新词的语义相似度；

将语义相似度大于第二预设阈值的词作为第二候选新词，并将所述第二候选新词添加到所述分词词典中；其中各个候选词的信息增益计算公式如下：

$$\mathrm{Gain}(t_i)=\left\{-\sum_{j=1}^{|c|}P(C_j)\times logP(C_j)\right\}-\left\{P(t_i)\times\left[-\sum_{j=1}^{|c|}P(C_j|t_i)\times logP(C_j|t_i)\right]+P(\overline{t_i})\times\left[-\sum_{j=1}^{|c|}P\left(C_j\middle|(\overline{t_i})\right)\times logP\left(C_j\middle|(\overline{t_i})\right)\right]\right\}$$

上述公式中，$P(C_j)$表示类别C_j在数据集中出现的概率，$P(t_i)$表示特征项t_i出现在数据集中的概率，$P(C_j|t_i)$表示特征项t_i出现在判定为类别C_j的文档中的概率，$P(\overline{t_i})$表示特征项t_i不出现的概率，$P\left(C_j\middle|(\overline{t_i})\right)$表示特征项$t_i$出现在不属于类别$C_j$的文档中的概率，$|c|$为类别的总数，其中类别是指情感倾向的分类，特征项是候选词，上述概率值都可以通过对候选词在文本语料中的统计情况计算得到；

获取样本集，按照预设情感倾向分类模式对所述样本集中的训练样本进行类别标注；

基于添加了候选新词的所述分词词典，使用预设的分词算法对所述样本集中的训练样本进行分词处理；

根据分词结果提取词向量，基于adaboost算法，将训练样本对应的词向量和标注的类别信息输入到预设的多个弱分类器中进行训练，将训练得到的多个弱分类器组合为金融领域的文本分类模型。

2.如权利要求1所述的文本分类模型的生成装置，其特征在于，所述处理器还可用于执行所述模型生成程序，以在所述将语义相似度大于第二预设阈值的词作为第二候选新词，并将所述第二候选新词添加到所述词词典的步骤之后，还实现如下步骤：

计算所述第二候选新词在文本语料中的词频，并将计算得到的词频作为该第二候选新词在所述分词词典中的权重。

3.如权利要求1或2中任一项所述的文本分类模型的生成装置，其特征在于，所述获取样本集，按照预设情感倾向分类模式对所述样本集中的训练样本进行类别标注的步骤包

括：

获取样本集，并获取多个标注人按照预设情感倾向分类模式对样本集中的训练样本进行标注得到的多个标注信息，从所述多个标注信息中，选择出现次数最多的标注信息作为对应的训练样本的标注结果。

4.如权利要求1或2中任一项所述的文本分类模型的生成装置，其特征在于，所述弱分类器包括基于卷积神经网络算法的分类器、基于循环神经网络算法的分类器和基于长短期记忆网络算法的分类器。

5.一种文本分类模型的生成方法，其特征在于，所述方法包括：

获取基于收集的金融领域词汇构建的金融领域的分词词典，以及预设的金融领域的文本语料；

根据预设算法从所述文本语料中选择候选新词，添加至所述分词词典，所述根据预设算法从所述文本语料中选择候选新词，添加至所述分词词典的步骤包括：

基于所述分词词典，使用所述分词算法对所述文本语料进行分词处理，根据所述分词结果获取候选词集合；

计算所述候选词集合中各个候选词的信息增益，选择信息增益大于第一预设阈值的候选词作为第一候选新词，将所述第一候选新词添加到所述分词词典中；

基于添加了所述第一候选新词的分词词典，使用所述分词算法对所述文本语料进行分词，并使用分词处理后的文本语料训练词向量模型；

使用训练得到的词向量模型计算分词结果中的词与所述第一候选新词的语义相似度；

将语义相似度大于第二预设阈值的词作为第二候选新词，并将所述第二候选新词添加到所述分词词典中；其中各个候选词的信息增益计算公式如下：

$$\mathrm{Gain}(t_i)=\left\{-\sum_{j=1}^{|c|}P(C_j)\times logP(C_j)\right\}-\left\{P(t_i)\times\left[-\sum_{j=1}^{|c|}P(C_j|t_i)\times logP(C_j|t_i)\right]+P(\overline{t_i})\times\left[-\sum_{j=1}^{|c|}P\left(C_j\middle|(\overline{t_i})\right)\times logP\left(C_j\middle|(\overline{t_i})\right)\right]\right\}$$

上述公式中，$P(C_j)$表示类别C_j在数据集中出现的概率，$P(t_i)$表示特征项t_i出现在数据集中的概率，$P(C_j|t_i)$表示特征项t_i出现在判定为类别C_j的文档中的概率，$P(\overline{t_i})$表示特征项t_i不出现的概率，$P\left(C_j\middle|(\overline{t_i})\right)$表示特征项$t_i$出现在不属于类别$C_j$的文档中的概率，$|c|$为类别的总数，其中类别是指情感倾向的分类，特征项是候选词，上述概率值都可以通过对候选词在文本语料中的统计情况计算得到；

获取样本集，按照预设情感倾向分类模式对所述样本集中的训练样本进行类别标注；

基于添加了候选新词的所述分词词典，使用预设的分词算法对所述样本集中的训练样本进行分词处理；

根据分词结果提取词向量，基于adaboost算法，将训练样本对应的词向量和标注的类别信息输入到预设的多个弱分类器中进行训练，将训练得到的多个弱分类器组合为金融领域的文本分类模型。

6.如权利要求5所述的文本分类模型的生成方法，其特征在于，所述将语义相似度大于第二预设阈值的词作为第二候选新词，并将所述第二候选新词添加到所述词词典的步骤之后，所述方法还包括步骤：

计算所述第二候选新词在文本语料中的词频，并将计算得到的词频作为该第二候选新词在所述分词词典中的权重。

7.如权利要求5或6中任一项所述的文本分类模型的生成方法，其特征在于，所述获取样本集，按照预设情感倾向分类模式对所述样本集中的训练样本进行类别标注的步骤包括：

获取样本集，并获取多个标注人按照预设情感倾向分类模式对样本集中的训练样本进行标注得到的多个标注信息，从所述多个标注信息中，选择出现次数最多的标注信息作为对应的训练样本的标注结果。

8.一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上存储有模型生成程序，所述模型生成程序可被一个或者多个处理器执行，以实现如权利要求5至7中任一项所述的文本分类模型的生成方法的步骤。

文本分类模型的生成装置、方法及计算机可读存储介质

技术领域

[0001] 本发明涉及文本分类技术领域，尤其涉及一种文本分类模型的生成装置、方法及计算机可读存储介质。

背景技术

[0002] 随着互联网和信息技术的发展，越来越多的机构和个人通过互联网途径以各种方式发表对各种事物的观点、态度和立场，如各种新闻评论、论坛以及社交网站等。这些海量的信息对于电子商务、市场预测等各个方面具有一定的商业价值，特别是金融行业，是互联网信息增长最快，受影响最大的行业，因此，对金融文本信息进行情感倾向分析以进行更加深入的研究逐渐成为重要课题。

[0003] 文本情感倾向性分析是属于文本情感分析的一部分，通过情感倾向性分析，可以掌握本文的褒贬性倾向，对于金融领域来说，新闻舆情是体现市场和行业的景气程度以及投资者的交易热情的重要指标，因此，对金融领域的文本的情感倾向性的分析对于金融时长的研究具有剧组轻重的影响，但是现有技术中还缺乏实现对金融领域文本进行情感倾向的分类的方案，导致无法实现对金融领域文本进行情感倾向性的分类。

发明内容

[0004] 本发明提供一种文本分类模型的生成装置、方法及计算机可读存储介质，其主要目的在于提出一种可以用于金融领域文本的情感倾向分类的文本分类模型的生成装置，以解决现有技术中无法实现对金融领域文本进行情感倾向性的分类的问题。

[0005] 为实现上述目的，本发明提供一种文本分类模型的生成装置，该装置包括存储器和处理器，所述存储器中存储有可在所述处理器上运行的模型生成程序，所述模型生成程序被所述处理器执行时实现如下步骤：

[0006] 获取基于收集的金融领域词汇构建的金融领域的分词词典，以及预设的金融领域的文本语料；

[0007] 根据预设算法从所述文本语料中选择候选新词，添加至所述分词词典；

[0008] 获取样本集，按照预设情感倾向分类模式对所述样本集中的训练样本进行类别标注；

[0009] 基于添加了候选新词的所述分词词典，使用预设的分词算法对所述样本集中的训练样本进行分词处理；

[0010] 根据分词结果提取词向量，基于adaboost算法，将训练样本对应的词向量和标注的类别信息输入到预设的多个弱分类器中进行训练，将训练得到的多个弱分类器组合为金融领域的文本分类模型。

[0011] 可选地，所述根据预设算法从所述文本语料中选择候选新词，添加至所述分词词典的步骤包括：

[0012] 基于所述分词词典，使用所述分词算法对所述文本语料进行分词处理，根据所述

分词结果获取候选词集合；

[0013] 计算所述候选词集合中各个候选词的信息增益，选择信息增益大于第一预设阈值的候选词作为第一候选新词，将所述第一候选新词添加到所述分词词典中；

[0014] 基于添加了所述第一候选新词的分词词典，使用所述分词算法对所述文本语料进行分词，并使用分词处理后的文本语料训练词向量模型；

[0015] 使用训练得到的词向量模型计算分词结果中的词与所述第一候选新词的语义相似度；

[0016] 将语义相似度大于第二预设阈值的词作为第二候选新词，并将所述第二候选新词添加到所述分词词典中。

[0017] 可选地，所述处理器还可用于执行所述模型生成程序，以在所述将语义相似度大于第二预设阈值的词作为第二候选新词，并将所述第二候选新词添加到所述词词典的步骤之后，还实现如下步骤：

[0018] 计算所述第二候选新词在文本语料中的词频，并将计算得到的词频作为该第二候选新词在所述分词词典中的权重。

[0019] 可选地，所述获取样本集，按照预设情感倾向分类模式对所述样本集中的训练样本进行类别标注的步骤包括：

[0020] 获取样本集，并获取多个标注人按照预设情感倾向分类模式对样本集中的训练样本进行标注得到的多个标注信息，从所述多个标注信息中，选择出现次数最多的标注信息作为对应的训练样本的标注结果。

[0021] 可选地，所述弱分类器包括基于卷积神经网络算法的分类器、基于循环神经网络算法的分类器和基于长短期记忆网络算法的分类器。

[0022] 此外，为实现上述目的，本发明还提供一种文本分类模型的生成方法，该方法包括：

[0023] 获取基于收集的金融领域词汇构建的金融领域的分词词典，以及预设的金融领域的文本语料；

[0024] 根据预设算法从所述文本语料中选择候选新词，添加至所述分词词典；

[0025] 获取样本集，按照预设情感倾向分类模式对所述样本集中的训练样本进行类别标注；

[0026] 基于添加了候选新词的所述分词词典，使用预设的分词算法对所述样本集中的训练样本进行分词处理；

[0027] 根据分词结果提取词向量，基于adaboost算法，将训练样本对应的词向量和标注的类别信息输入到预设的多个弱分类器中进行训练，将训练得到的多个弱分类器组合为金融领域的文本分类模型。

[0028] 可选地，所述根据预设算法从所述文本语料中选择候选新词，添加至所述分词词典的步骤包括：

[0029] 基于所述分词词典，使用所述分词算法对所述文本语料进行分词处理，根据所述分词结果获取候选词集合；

[0030] 计算所述候选词集合中各个候选词的信息增益，选择信息增益大于第一预设阈值的候选词作为第一候选新词，将所述第一候选新词添加到所述分词词典中；

[0031] 基于添加了所述第一候选新词的分词词典，使用所述分词算法对所述文本语料进行分词，并使用分词处理后的文本语料训练词向量模型；

[0032] 使用训练得到的词向量模型计算分词结果中的词与所述第一候选新词的语义相似度；

[0033] 将语义相似度大于第二预设阈值的词作为第二候选新词，并将所述第二候选新词添加到所述分词词典中。

[0034] 可选地，所述将语义相似度大于第二预设阈值的词作为第二候选新词，并将所述第二候选新词添加到所述词词典的步骤之后，所述方法还包括步骤：

[0035] 计算所述第二候选新词在文本语料中的词频，并将计算得到的词频作为该第二候选新词在所述分词词典中的权重。

[0036] 可选地，所述获取样本集，按照预设情感倾向分类模式对所述样本集中的训练样本进行类别标注的步骤包括：

[0037] 获取样本集，并获取多个标注人按照预设情感倾向分类模式对样本集中的训练样本进行标注得到的多个标注信息，从所述多个标注信息中，选择出现次数最多的标注信息作为对应的训练样本的标注结果。

[0038] 此外，为实现上述目的，本发明还提供一种计算机可读存储介质，所述计算机可读存储介质上存储有模型生成程序，所述模型生成程序可被一个或者多个处理器执行，以实现如上所述的文本分类模型的生成方法的步骤。

[0039] 本发明提出的文本分类模型的生成装置、方法及计算机可读存储介质，基于收集的金融领域词汇构建金融领域的分词词典，获取预设的金融领域的文本语料，根据文本语料获取候选词集合，从候选词集合中选择候选新词添加至分词词典。获取样本集，按照预设情感倾向分类模式对样本集中的训练样本进行类别标注。基于添加了候选新词的分词词典，使用预设的分词算法对样本集中的训练样本进行分词处理，根据分词结果提取词向量，基于adaboost算法，将训练样本对应的词向量和标注的类别信息输入到预设的多个弱分类器中进行训练，将训练得到的多个弱分类器组合为金融领域的文本分类模型。本发明的方案中，通过对金融领域的文本语料挖掘，筛选出候选新词添加到分词词典中，通过更新后的分词词典对样本集中的样本分词处理，并按照预设情感倾向分类模式对样本集中的样本数据进行类别标注，最终训练得到文本分类模型，该模型能够应用于金融领域的情感倾向分类问题。

附图说明

[0040] 图1为本发明文本分类模型的生成装置较佳实施例的示意图；

[0041] 图2为本发明文本分类模型的生成装置一实施例中模型生成程序的程序模块示意图；

[0042] 图3为本发明文本分类模型的生成方法较佳实施例的流程图。

[0043] 本发明目的的实现、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

[0044] 应当理解，此处所描述的具体实施例仅仅用以解释本发明，并不用于限定本发明。

[0045] 本发明提供一种文本分类模型的生成装置。参照图1所示，为本发明文本分类模型的生成装置较佳实施例的示意图。

[0046] 在本实施例中，文本分类模型的生成装置可以是PC(Personal Computer，个人电脑)，也可以是智能手机、平板电脑、便携计算机等终端设备。该文本分类模型的生成装置1至少包括存储器11、处理器12，通信总线13，以及网络接口14。

[0047] 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文本分类模型的生成装置1的内部存储单元，例如该文本分类模型的生成装置1的硬盘。存储器11在另一些实施例中也可以是文本分类模型的生成装置1的外部存储设备，例如文本分类模型的生成装置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还可以既包括文本分类模型的生成装置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文本分类模型的生成装置1的应用软件及各类数据，例如模型生成程序01的代码等，还可以用于暂时地存储已经输出或者将要输出的数据。

[0048] 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，例如执行模型生成程序01等。

[0049] 通信总线13用于实现这些组件之间的连接通信。

[0050] 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)，通常用于在该装置1与其他电子设备之间生成通信连接。

[0051] 图1仅示出了具有组件11-14以及模型生成程序01的文本分类模型的生成装置1，但是应理解的是，并不要求实施所有示出的组件，可以替代的实施更多或者更少的组件。

[0052] 可选地，该装置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。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OLED(Organic Light-Emitting Diode，有机发光二极管)触摸器等。其中，显示器也可以适当的称为显示屏或显示单元，用于显示在文本分类模型的生成装置1中处理的信息以及用于显示可视化的用户界面。

[0053] 在图1所示的装置1实施例中，存储器11中存储有模型生成程序01；处理器12执行存储器11中存储的模型生成程序01时实现如下步骤：

[0054] A1、获取基于收集的金融领域词汇构建的金融领域的分词词典，以及预设的金融领域的文本语料。

[0055] 首先，获取全领域分词词典，在全领域分词词典的基础上，加入收集的金融领域的词汇构成金融领域分词词典。其中，金融领域的词汇来源主要包括以下三类：金融领域专业术语，例如“威廉指标”、“移动平均线”、“可转债”等；金融论坛用语，如一些炒股论坛中的用户在评论股票时的用语；应用于金融领域的网络用语和特定符号，例如“垃圾股”等。

[0056] A2、根据预设算法从所述文本语料中选择候选新词，添加至所述分词词典。

[0057] 在上述分词词典的基础上，从新的文本预料中选择候选新词添加到当前的分词词典中。具体地，步骤A2包括：

[0058] A21、基于所述分词词典，使用所述分词算法对所述文本语料进行分词处理，根据所述分词结果获取候选词集合；A22、计算所述候选词集合中各个候选词的信息增益，选择信息增益大于第一预设阈值的候选词作为第一候选新词，将所述第一候选新词添加到所述分词词典中；A23、基于添加了所述第一候选新词的分词词典，使用所述分词算法对所述文本语料进行分词，并使用分词处理后的文本语料训练词向量模型；A24、使用训练得到的词向量模型计算分词结果中的词与所述第一候选新词的语义相似度；A25、将语义相似度大于第二预设阈值的词作为第二候选新词，并将所述第二候选新词添加到所述分词词典中。

[0059] 获取用于扩充分词词典的文本语料。具体地，采用网络爬虫的方式从金融网站抓取大量的、与待分析的金融主题相关的金融新闻文本信息，形成文本语料。对爬取到的数据进行预处理，去除其中包含的乱码符号、web转义符号等无用信息，保留文本数据作为文本语料。接下来，通过人工标注的方式对文本语料中的大量文本数据进行情感倾向的分类，即为文本数据添加类别标注信息。

[0060] 以当前的分词词典作为预设的分词算法的词典，对文本语料进行分词处理，然后，根据预设的停用词词表过滤掉分词结果中的停用词，以去掉结果中的无关词汇，由剩余的分词结果构成候选词集合。分词结果对应的类别标注信息与其对应的文本数据的类别标注信息一致。

[0061] 接下来，根据信息增益从候选词集合中选择候选新词，其中，信息增益是一种基于熵的评估方法，其用于特征选择时，衡量的是某个词的出现与否对判断一个文本是否属于某个类所提供的信息量；其定义为某一特征值在文档中出现前后的信息量之差，计算公式为：

[0062] $$\mathrm{Gain}(t_i)=\left\{-\sum_{j=1}^{|c|}P(C_j)\times logP(C_j)\right\}-\left\{P(t_i)\times\left[-\sum_{j=1}^{|c|}P(C_j|t_i)\times logPCjti+Pti\times-j=1cPCjti\times logPCjti\right.\right.$$

[0063] 上述公式中，$P(C_j)$表示类别C_j在数据集中出现的概率，$P(t_i)$表示特征项t_i出现在数据集中的概率，$P(C_j|t_i)$表示特征项t_i出现在判定为类别C_j的文档中的概率，$P(\overline{t_i})$表示特征项t_i不出现的概率，$P\left(C_j\middle|(\overline{t_i})\right)$表示特征项$t_i$出现在不属于类别$C_j$的文档中的概率，$|c|$为类别的总数。其中，类别是指情感倾向的分类，特征项是候选词。上述概率值都可以通过对候选词在文本语料中的统计情况计算得到。

[0064] 根据计算得到的信息增益判断候选词的有用程度，信息增益的值越大，则对分类越有用。将候选词集合中信息增益大于第一预设阈值的候选词作为第一候选新词，添加到当前的分词词典中，实现对分词词典的扩充。

[0065] 基于上述经过词汇扩充的分词词典，使用同样的分词算法对上述同样的文本语料进行分词处理，获取分词结果，使用分词处理后的文本语料训练词向量模型，使用训练得到的词向量模型计算分词得到的各个词的词向量，根据词向量计算分词处理得到的词语第一候选新词的语义相似度，若语义相似度大于第二预设阈值，则将其作为第二候选新词，将从分词结果中选择出的第二候选词添加到分词词典中，实现对分词词典的再次扩充。

[0066] 可以理解的是，在使用分词算法对文本语料进行分词处理后，会通过停用词词表删除掉分词结果中的停用词，因为这些停用词噪音大，且对文本分类无意义，删除这些词可

以提高文本分类的准确程度，同时减小选择候选新词时的计算量。

[0067] 经过上述步骤，实际上实现了对分词词典的三次扩充，第一次是通过人工收集的方式获取金融领域词汇的方式进行初步的扩充，第二次是通过计算信息增益选择新词，第三次是通过词向量计算语义相似度再次选择新词。此外，第二次和第三次扩充均是在上一次扩充后的分词词典的基础上进行的再次扩充。通过上述动态扩充的方式，尽可能多地从语料中筛选新的金融领域词汇。扩充了分词词典的分词算法用于分类模型的训练样本的分词，分词词典中的金融领域词汇越丰富，则对金融领域文本的分词结果越准确，训练得到的分类模型的分类准确度也越高。

[0068] 可选地，作为一种实施方式，在选择了第二候选新词并添加至分词词典中之后，计算第二候选新词在文本语料中的词频，并且将词频作为第二候选新词在分词词典中的权重。对于第一候选新词可以采用同样的方式计算词频并作为其在分词词典中的权重。

[0069] A3、获取样本集，按照预设情感倾向分类模式对所述样本集中的训练样本进行类别标注。

[0070] 获取用于训练文本分类模型的样本集，针对样本集中的每一条训练数据，获取多个标注人按照预设情感倾向分类模式对每一条训练数据的多个标注信息，并选择多个标注信息中出现次数最多的标注信息作为该训练数据的标注结果。其中，用户可以根据待分析的金融问题设置对应的情感倾向分类模式，例如，将股票论坛中的文本划分为持有、卖出和买入；将微博或者论坛中的股票讨论文本划分为积极、消极和中性；将财经新闻文本划分为正向、负向和中性等。

[0071] A4、基于添加了候选新词的所述分词词典，使用预设的分词算法对所述样本集中的训练样本进行分词处理。

[0072] A5、根据分词结果提取词向量，基于adaboost算法，将训练样本对应的词向量和标注的类别信息输入到预设的多个弱分类器中进行训练，将训练得到的多个弱分类器组合为金融领域的文本分类模型。

[0073] 在完成了对训练样本的标注之后，基于经过多次扩充的分词词典，使用预设分词算法对于训练样本分词处理，使用训练后的词向量模型提取分词结果的词向量。需要说明的是，本实施例的方案中采用的分词算法始终为同一个算法。

[0074] 在本实施例中，为了提高文本分类模型的准确度，分别使用word2vec模型和Glove(Global Vectors for word representation)模型提取词向量，每一个分词结果都会得到两种词向量。此外，本实施例中将基于卷积神经网络算法的分类器、基于循环神经网络算法的分类器和基于长短期记忆网络算法的分类器作为弱分类器。针对每一个弱分类器，分别将上述两种词向量作为输入，则实际上可以构建六个弱分类模型。基于adaboost算法，使用样本集中的样本训练各个弱分类器。在训练过程中，如果某样本已经被准确地分类，那么在构造下一个样本集中，降低该样本的权值；如果某样本未被准确地分类，那么在构造下一个样本集中，提高该样本的权值。权值更新过的样本集被用于训练下一个分类器，整个训练过程如此迭代地进行下去。此外，各个弱分类器的训练过程结束后，加大分类误差率小的弱分类器的权重，使其在最终的分类函数中起着较大的决定作用，而降低分类误差率大的弱分类器的权重，使其在最终的分类函数中起着较小的决定作用。按照上述过程迭代地训练各个弱分类器。将每次训练得到的弱分类器融合起来，作为最终的文本分类模型。该文本分类

模型可以用于对金融领域文本进行情感倾向的分类，用于判断论坛中的股票讨论文本是消极、积极或者中性等。

[0075] 本实施例提出的文本分类模型的生成装置，通过对金融领域的文本语料挖掘，尽可能多的从语料中筛选新的金融领域词语，添加到分词词典中，实现对金融领域分词词典的扩充，并且使用扩充了金融词汇后的分词词典对样本集中的训练样本进行分词处理，并按照预设情感倾向分类模式对样本集中的样本数据进行类别标注，最终训练得到文本分类模型，该模型能够应用于金融领域的情感倾向分类问题。

[0076] 可选地，在其他的实施例中，模型生成程序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)所执行以完成本发明，本发明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，用于描述模型生成程序在文本分类模型的生成装置中的执行过程。

[0077] 例如，参照图2所示，为本发明文本分类模型的生成装置一实施例中的模型生成程序的程序模块示意图，该实施例中，模型生成程序可以被分割为数据获取模块10、新词选择模块20、样本标注模块30、样本分词模块40和模型训练模块50，示例性地：

[0078] 数据获取模块10用于：获取基于收集的金融领域词汇构建的金融领域的分词词典，以及预设的金融领域的文本语料；

[0079] 新词选择模块20用于：根据预设算法从所述文本语料中选择候选新词，添加至所述分词词典；

[0080] 样本标注模块30用于：获取样本集，按照预设情感倾向分类模式对所述样本集中的训练样本进行类别标注；

[0081] 样本分词模块40用于：基于添加了候选新词的所述分词词典，使用预设的分词算法对所述样本集中的训练样本进行分词处理；

[0082] 模型训练模块50用于：根据分词结果提取词向量，基于adaboost算法，将训练样本对应的词向量和标注的类别信息输入到预设的多个弱分类器中进行训练，将训练得到的多个弱分类器组合为金融领域的文本分类模型。

[0083] 上述数据获取模块10、新词选择模块20、样本标注模块30、样本分词模块40和模型训练模块50等程序模块被执行时所实现的功能或操作步骤与上述实施例大体相同，在此不再赘述。

[0084] 此外，本发明还提供一种文本分类模型的生成方法。参照图3所示，为本发明文本分类模型的生成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。

[0085] 在本实施例中，文本分类模型的生成方法包括：

[0086] 步骤S10，获取基于收集的金融领域词汇构建的金融领域的分词词典，以及预设的金融领域的文本语料。

[0087] 首先，获取全领域分词词典，在全领域分词词典的基础上，加入收集的金融领域的词汇构成金融领域分词词典。其中，金融领域的词汇来源主要包括以下三类：金融领域专业术语，例如“威廉指标”、“移动平均线”、“可转债”等；金融论坛用语，如一些炒股论坛中的用户在评论股票时的用语；应用于金融领域的网络用语和特定符号，例如“垃圾股”等。

[0088] 步骤S20，根据预设算法从所述文本语料中选择候选新词，添加至所述分词词典。

[0089] 在上述分词词典的基础上，从新的文本预料中选择候选新词添加到当前的分词词典中。具体地，步骤S20包括：基于所述分词词典，使用所述分词算法对所述文本语料进行分词处理，根据所述分词结果获取候选词集合；计算所述候选词集合中各个候选词的信息增益，选择信息增益大于第一预设阈值的候选词作为第一候选新词，将所述第一候选新词添加到所述分词词典中；基于添加了所述第一候选新词的分词词典，使用所述分词算法对所述文本语料进行分词，并使用分词处理后的文本语料训练词向量模型；使用训练得到的词向量模型计算分词结果中的词与所述第一候选新词的语义相似度；将语义相似度大于第二预设阈值的词作为第二候选新词，并将所述第二候选新词添加到所述分词词典中。

[0090] 获取用于扩充分词词典的文本语料。具体地，采用网络爬虫的方式从金融网站抓取大量的、与待分析的金融主题相关的金融新闻文本信息，形成文本语料。对爬取到的数据进行预处理，去除其中包含的乱码符号、web转义符号等无用信息，保留文本数据作为文本语料。接下来，通过人工标注的方式对文本语料中的大量文本数据进行情感倾向的分类，即为文本数据添加类别标注信息。

[0091] 以当前的分词词典作为预设的分词算法的词典，对文本语料进行分词处理，然后，根据预设的停用词词表过滤掉分词结果中的停用词，以去掉结果中的无关词汇，由剩余的分词结果构成候选词集合。分词结果对应的类别标注信息与其对应的文本数据的类别标注信息一致。

[0092] 接下来，根据信息增益从候选词集合中选择候选新词，其中，信息增益是一种基于熵的评估方法，其用于特征选择时，衡量的是某个词的出现与否对判断一个文本是否属于某个类所提供的信息量；其定义为某一特征值在文档中出现前后的信息量之差，计算公式为：

[0093] $$\mathrm{Gain}(t_i)=\left\{-\sum_{j=1}^{|c|}P(C_j)\times logP(C_j)\right\}-\left\{P(t_i)\times\left[-\sum_{j=1}^{|c|}P(C_j|t_i)\times logPCjti+Pti\times-j=1cPCjti\times logPCjti\right.\right.$$

[0094] 上述公式中，$P(C_j)$表示类别C_j在数据集中出现的概率，$P(t_i)$表示特征项t_i出现在数据集中的概率，$P(C_j|t_i)$表示特征项t_i出现在判定为类别C_j的文档中的概率，$P(\overline{t_i})$表示特征项t_i不出现的概率，$P\left(C_j\middle|(\overline{t_i})\right)$表示特征项$t_i$出现在不属于类别$C_j$的文档中的概率，$|c|$为类别的总数。其中，类别是指情感倾向的分类，特征项是候选词。上述概率值都可以通过对候选词在文本语料中的统计情况计算得到。

[0095] 根据计算得到的信息增益判断候选词的有用程度，信息增益的值越大，则对分类越有用。将候选词集合中信息增益大于第一预设阈值的候选词作为第一候选新词，添加到当前的分词词典中，实现对分词词典的扩充。

[0096] 基于上述经过词汇扩充的分词词典，使用同样的分词算法对上述同样的文本语料进行分词处理，获取分词结果，使用分词处理后的文本语料训练词向量模型，使用训练得到的词向量模型计算分词得到的各个词的词向量，根据词向量计算分词处理得到的词语第一候选新词的语义相似度，若语义相似度大于第二预设阈值，则将其作为第二候选新词，将从分词结果中选择出的第二候选词添加到分词词典中，实现对分词词典的再次扩充。

[0097] 可以理解的是，在使用分词算法对文本语料进行分词处理后，会通过停用词词表

删除掉分词结果中的停用词，因为这些停用词噪音大，且对文本分类无意义，删除这些词可以提高文本分类的准确程度，同时减小选择候选新词时的计算量。

[0098] 经过上述步骤，实际上实现了对分词词典的三次扩充，第一次是通过人工收集的方式获取金融领域词汇的方式进行初步的扩充，第二次是通过计算信息增益选择新词，第三次是通过词向量计算语义相似度再次选择新词。此外，第二次和第三次扩充均是在上一次扩充后的分词词典的基础上进行的再次扩充。通过上述动态扩充的方式，尽可能多地从语料中筛选新的金融领域词汇。扩充了分词词典的分词算法用于分类模型的训练样本的分词，分词词典中的金融领域词汇越丰富，则对金融领域文本的分词结果越准确，训练得到的分类模型的分类准确度也越高。

[0099] 可选地，作为一种实施方式，在选择了第二候选新词并添加至分词词典中之后，计算第二候选新词在文本语料中的词频，并且将词频作为第二候选新词在分词词典中的权重。对于第一候选新词可以采用同样的方式计算词频并作为其在分词词典中的权重。

[0100] 步骤S30，获取样本集，按照预设情感倾向分类模式对所述样本集中的训练样本进行类别标注。

[0101] 获取用于训练文本分类模型的样本集，针对样本集中的每一条训练数据，获取多个标注人按照预设情感倾向分类模式对每一条训练数据的多个标注信息，并选择多个标注信息中出现次数最多的标注信息作为该训练数据的标注结果。其中，用户可以根据待分析的金融问题设置对应的情感倾向分类模式，例如，将股票论坛文本划分为持有、卖出和买入；将微博股票讨论文本划分为积极、消极和中性；将财经新闻文本划分为正向、负向和中性等。

[0102] 步骤S40，基于添加了候选新词的所述分词词典，使用预设的分词算法对所述样本集中的训练样本进行分词处理。

[0103] 步骤S50，根据分词结果提取词向量，基于adaboost算法，将训练样本对应的词向量和标注的类别信息输入到预设的多个弱分类器中进行训练，将训练得到的多个弱分类器组合为金融领域的文本分类模型。

[0104] 在完成了对训练样本的标注之后，基于经过多次扩充的分词词典，使用预设分词算法对于训练样本分词处理，使用训练后的词向量模型提取分词结果的词向量。需要说明的是，本实施例的方案中采用的分词算法始终为同一个算法。

[0105] 在本实施例中，为了提高文本分类模型的准确度，分别使用word2vec模型和Glove模型提取词向量，每一个分词结果都会得到两种词向量。此外，本实施例中将基于卷积神经网络算法的分类器、基于循环神经网络算法的分类器和基于长短期记忆网络算法的分类器作为弱分类器。针对每一个弱分类器，分别将上述两种词向量作为输入，则实际上可以构建六个弱分类模型。基于adaboost算法，使用样本集中的样本训练各个弱分类器。在训练过程中，如果某样本已经被准确地分类，那么在构造下一个样本集中，降低该样本的权值；如果某样本未被准确地分类，那么在构造下一个样本集中，提高该样本的权值。权值更新过的样本集被用于训练下一个分类器，整个训练过程如此迭代地进行下去。此外，各个弱分类器的训练过程结束后，加大分类误差率小的弱分类器的权重，使其在最终的分类函数中起着较大的决定作用，而降低分类误差率大的弱分类器的权重，使其在最终的分类函数中起着较小的决定作用。按照上述过程迭代地训练各个弱分类器。将每次训练得到的弱分类器融合

起来，作为最终的文本分类模型。该文本分类模型可以用于对金融领域文本进行情感倾向的分类，用于判断论坛中的股票讨论文本是消极、积极或者中性等。

[0106] 本实施例提出的文本分类模型的生成方法，通过对金融领域的文本语料挖掘，尽可能多的从语料中筛选新的金融领域词语，添加到分词词典中，实现对金融领域分词词典的扩充，并且使用扩充了金融词汇后的分词词典对样本集中的训练样本进行分词处理，并按照预设情感倾向分类模式对样本集中的样本数据进行类别标注，最终训练得到文本分类模型，该模型能够应用于金融领域的情感倾向分类问题。

[0107] 此外，本发明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上存储有模型生成程序，所述模型生成程序可被一个或多个处理器执行，以实现如下操作：

[0108] 获取基于收集的金融领域词汇构建的金融领域的分词词典，以及预设的金融领域的文本语料；

[0109] 根据预设算法从所述文本语料中选择候选新词，添加至所述分词词典；

[0110] 获取样本集，按照预设情感倾向分类模式对所述样本集中的训练样本进行类别标注；

[0111] 基于添加了候选新词的所述分词词典，使用预设的分词算法对所述样本集中的训练样本进行分词处理；

[0112] 根据分词结果提取词向量，基于adaboost算法，将训练样本对应的词向量和标注的类别信息输入到预设的多个弱分类器中进行训练，将训练得到的多个弱分类器组合为金融领域的文本分类模型。

[0113] 本发明计算机可读存储介质具体实施方式与上述文本分类模型的生成装置和方法各实施例基本相同，在此不作累述。

[0114] 需要说明的是，上述本发明实施例序号仅仅为了描述，不代表实施例的优劣。并且本文中的术语“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。

[0115] 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)执行本发明各个实施例所述的方法。

[0116] 以上仅为本发明的优选实施例，并非因此限制本发明的专利范围，凡是利用本发明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他相关的技术领域，均同理包括在本发明的专利保护范围内。

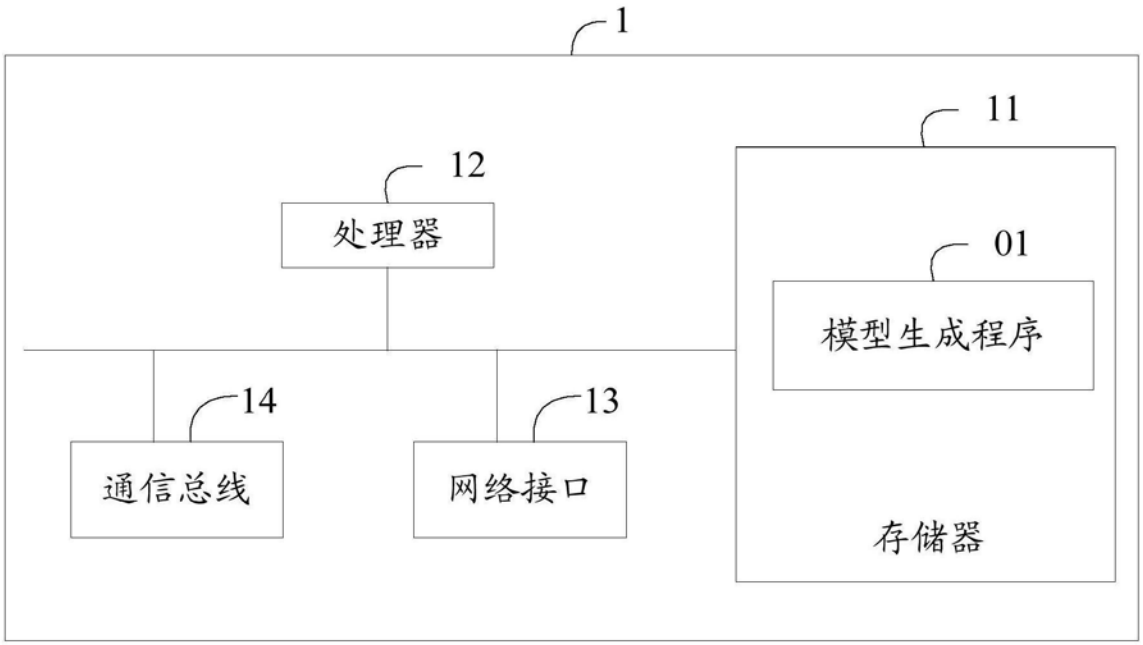

1
11
12
处理器
01
模型生成程序
14
13
通信总线
网络接口
存储器

图1

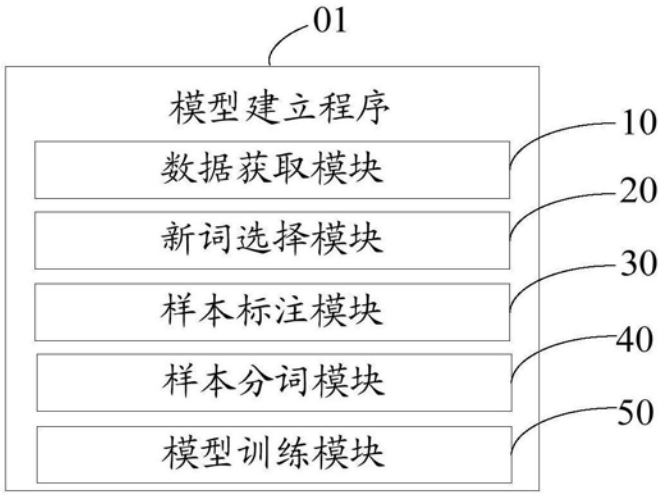

01
模型建立程序
10
数据获取模块
20
新词选择模块
30
样本标注模块
40
样本分词模块
50
模型训练模块

图2

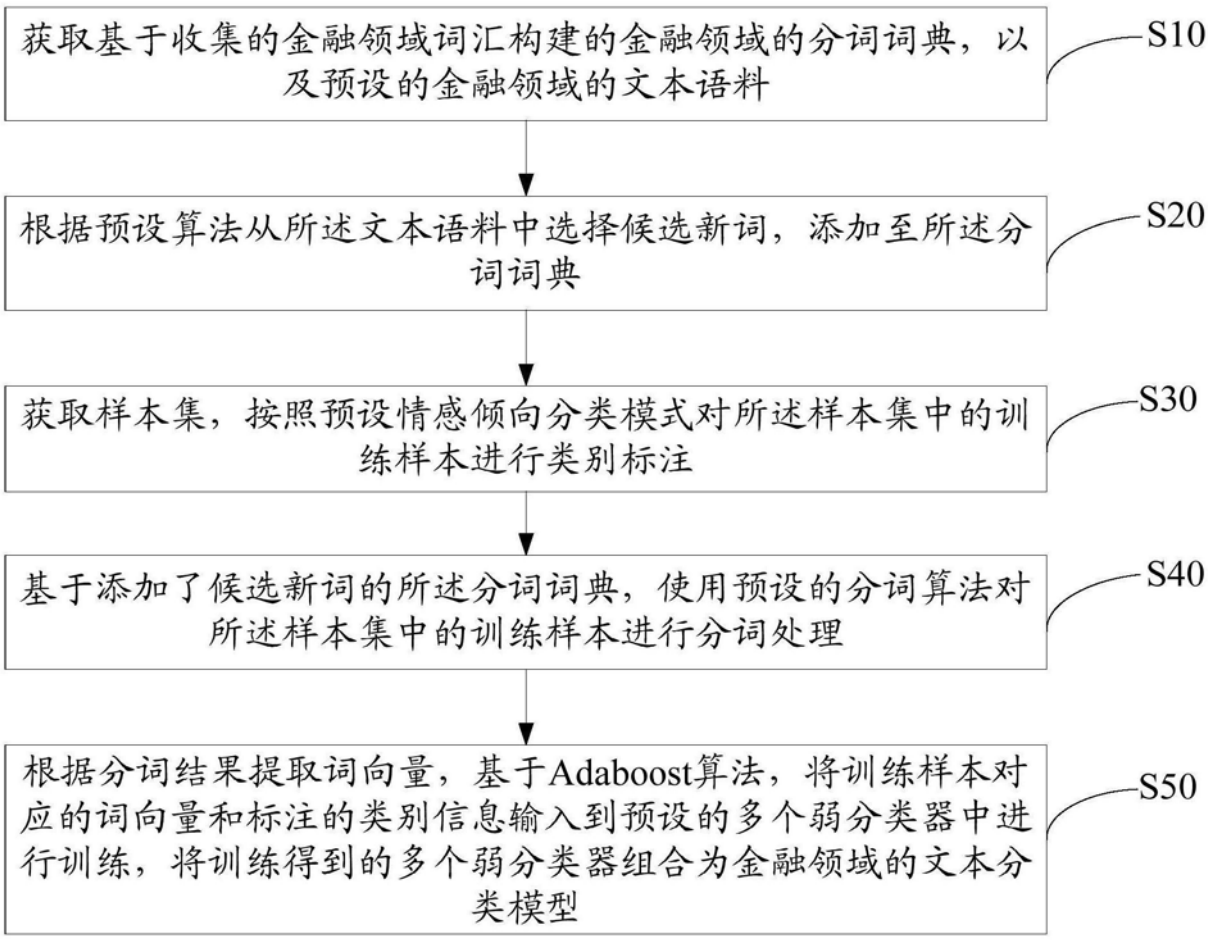

获取基于收集的金融领域词汇构建的金融领域的分词词典，以及预设的金融领域的文本语料
S10
根据预设算法从所述文本语料中选择候选新词，添加至所述分词词典
S20
获取样本集，按照预设情感倾向分类模式对所述样本集中的训练样本进行类别标注
S30
基于添加了候选新词的所述分词词典，使用预设的分词算法对所述样本集中的训练样本进行分词处理
S40
根据分词结果提取词向量，基于Adaboost算法，将训练样本对应的词向量和标注的类别信息输入到预设的多个弱分类器中进行训练，将训练得到的多个弱分类器组合为金融领域的文本分类模型
S50

图3