**权 利 要 求 书**

|  |
| --- |
|  |

1. 一种基于舆论数据的股票异动分析方法，包括如下步骤：

（1）对舆论数据进行采集并通过预处理得到多个样本集，所述样本集为固定时间段内关于股票的所有文章及其评论集合，其中的样本即为关于任一股票的一篇文章及其评论；

（2）根据股票龙虎榜中的股票异动类型对样本进行类别标注，并将具有类别标注信息的样本分为训练集、验证集和测试集；

（3）获取金融领域词典、情感词典以及程度副词词典，根据上述词典或通过word2vec方法对样本进行分词，即将样本中的每个词转换成词向量形式，从而得到每个句子由词向量所组成的词向量矩阵；

（4）将一个样本中长度不同的句子所对应的词向量矩阵统一转换成固定维度大小的特征向量，进而将所有句子对应的特征向量合并后作为样本的特征向量；

（5）将同一天内关于同一股票的所有样本的特征向量级联拼接成一个特征向量X，进而将所有特征向量X输入至CNN中，CNN的输出为对应一个日期戳的特征向量Y；

（6）对于任一发生异动的股票，将该股票异动发生前一定时间段内的所有特征向量Y级联拼接成一个特征向量Z，进而将所有特征向量Z输入至LSTM中，LSTM的输出为对应一种异动类型的特征向量W；

（7）将特征向量W、关于单股发帖量的特征信息以及大盘发帖量的变化特征信息共同输入至SVM中进行训练，得到用于股票异动分析的模型，进而利用该模型对股票异动提前做出分析预判。

2. 根据权利要求1所述的股票异动分析方法，其特征在于：所述步骤（1）中通过爬虫采集得到大量舆论数据，在对这些舆论数据做预处理时，根据模型中定义的句子长度的取值范围去除异常值或将超出范围值整理到正常范围中，根据文本的长度分布信息对舆论数据进行特定的预处理，预处理后即得到规整的舆论数据。

3. 根据权利要求1所述的股票异动分析方法，其特征在于：所述步骤（2）中根据最近一段时间的龙虎榜，对于其中存在异动的任一股票，获取该股票的异动类型及其异动发生时间，从样本集中查找异动发生时间之前的一定时间段内对应该股票的所有样本，并以该股票的异动类型对这些样本进行类别标注。

4. 根据权利要求1所述的股票异动分析方法，其特征在于：所述步骤（3）中对样本的结构化表示，即用数字向量的形式有效地表达一个非结构化的句子；方案一中对于每个句子，根据语义把句子中的词语转化为高维的词向量，将语义相近的词语置于相近的位置；方案二则基于金融领域词典、情感词典以及程度副词词典将句子中的词语表示为数据向量。

5. 根据权利要求1所述的股票异动分析方法，其特征在于：所述步骤（4）中采用RNN算法将一个样本中长度不同的句子所对应的词向量矩阵统一转换成固定维度大小的特征向量。

6. 根据权利要求1所述的股票异动分析方法，其特征在于：所述步骤（7）中得到用于股票异动分析的模型后，利用验证集对该模型进行性能评估及调试，最后将测试集样本输入至调试后最终的模型中进行测试，对应生成关于股票异动类型的分类结果。

**说 明 书**

|  |
| --- |
|  |

**一种基于舆论数据的股票异动分析方法**

**技术领域**

本发明属于自然语言处理及数据挖掘技术领域，具体涉及一种基于舆论数据的股票异动分析方法。

**背景技术**

随着互联网技术的飞速发展，信息化与金融行业的深度融合，专家和权威机构在互联网上发布的海量有价值的评论、重要金融事件的报道和公司披露等，已成为领域专家和投资者了解金融市场的重要的信息来源。然而，与股票行情走势和投资相关性较强的关键信息被隐藏在庞大的文本中，另人眼花缭乱，不能及时地获取有效信息。

行为经济学表明情绪会深刻地影响个人的行为和决策，这是否也适用于股票行业，即股民的情绪状态可能对其交易行为产生影响；股民的情感信息是否与股票的行情走势相关，甚至能否基于异常波动趋势之前的相关文本内容预测股票的异常波动？因此，基于趋势之前的舆论数据对股票异常波动进行分析具有相当重要的价值意义。

现阶段的金融数据分析，主要是金融专家基于相关的理论基础，对于海量的金融信息报道，结合国家的宏观经济政策、投资环境与理念、上市公司及其产品服务等因素，对股市的行情进行全面的分析，对股市可能发生的异常波动作出分析和预警。但由于股票相关文本的数量庞大、实时性强，人工逐条阅读分析需要消耗巨大的人力资源，需要金融分析人员长时间、高频率地进行文本阅读，不仅效率低下，而且由于工作状态、个体情绪等人为因素，会对异动分析的准确度造成影响。显然，如果能够对股票文本的情感特征进行自动提取，挖掘情感特征与股票异常波动的关系，无疑将会对证券行业和股民的投资行为给予非常有意义的帮助。

**发明内容**

鉴于上述，本发明提供了一种基于舆论数据的股票异动分析方法，其建立了一套文本情感分析的模型，通过自然语言处理技术，提取海量的舆论数据的特征，挖掘出与特定的股票异常波动类型的特征，帮助领域专家和数据分析师分析股票的行情，帮助投资者获取关键信息，改善投资策略，并给定通用的方法，大大提高金融文本数据分析的效率，符合领域专家及投资者的需要。

一种基于舆论数据的股票异动分析方法，包括如下步骤：

（1）对舆论数据进行采集并通过预处理得到多个样本集，所述样本集为固定时间段内关于股票的所有文章及其评论集合，其中的样本即为关于任一股票的一篇文章及其评论；

（2）根据股票龙虎榜中的股票异动类型对样本进行类别标注，并将具有类别标注信息的样本分为训练集、验证集和测试集；

（3）获取金融领域词典、情感词典以及程度副词词典，根据上述词典或通过word2vec方法对样本进行分词，即将样本中的每个词转换成词向量形式，从而得到每个句子由词向量所组成的词向量矩阵；

（4）将一个样本中长度不同的句子所对应的词向量矩阵统一转换成固定维度大小的特征向量，进而将所有句子对应的特征向量合并后作为样本的特征向量；

（5）将同一天内关于同一股票的所有样本的特征向量级联拼接成一个特征向量X，进而将所有特征向量X输入至CNN(Convolutional Neural Networks，卷积神经网络)中，CNN的输出为对应一个日期戳(即一天)的特征向量Y；

（6）对于任一发生异动的股票，将该股票异动发生前一定时间段内的所有特征向量Y级联拼接成一个特征向量Z，进而将所有特征向量Z输入至LSTM(Long Short-Term Memory，长短期记忆网络)中，LSTM的输出为对应一种异动类型的特征向量W；

（7）将特征向量W、关于单股发帖量的特征信息以及大盘发帖量的变化特征信息共同输入至SVM(Support Vector Machine，支持向量机)中进行训练，得到用于股票异动分析的模型，进而利用该模型对股票异动提前做出分析预判。

进一步地，所述步骤（1）中通过爬虫采集得到大量舆论数据，在对这些舆论数据做预处理时，根据模型中定义的句子长度的取值范围去除异常值或将超出范围值整理到正常范围中，根据文本的长度分布信息对舆论数据进行特定的预处理，预处理后即得到规整的舆论数据。

进一步地，所述步骤（2）中根据最近一段时间的龙虎榜，对于其中存在异动的任一股票，获取该股票的异动类型及其异动发生时间，从样本集中查找异动发生时间之前的一定时间段内对应该股票的所有样本，并以该股票的异动类型对这些样本进行类别标注。

进一步地，所述步骤（3）中对样本的结构化表示，即用数字向量的形式有效地表达一个非结构化的句子；方案一(即word2vec方法)中对于每个句子，根据语义把句子中的词语转化为高维的词向量，将语义相近的词语置于相近的位置；方案二则基于金融领域词典、情感词典以及程度副词词典将句子中的词语表示为数据向量。

进一步地，所述步骤（4）中采用RNN(Recurrent Neural Networks，循环神经网络)算法将一个样本中长度不同的句子所对应的词向量矩阵统一转换成固定维度大小的特征向量。

进一步地，所述步骤（7）中得到用于股票异动分析的模型后，利用验证集对该模型进行性能评估及调试，最后将测试集样本输入至调试后最终的模型中进行测试，对应生成关于股票异动类型的分类结果。

基于上述技术方案，本发明具有以下有益技术效果：

(1) 本发明给出了一个有效的文本分布式表示方法，对于不同文本的处理过程均适用。

(2) 本发明提供了一种提取舆论数据中与股票异动有关的特征的方法，能有效地提取异动发生前的特征，当出现包含这些特征的舆论时，领域专家可以结合其他特征，对股票可能出现的异常波动作出分析和及时的预警，使得领域专家在股票异动分析的工作中尽量发挥其领域知识方面的优势。

(3) 本发明提供了一个有效的文本特征提取模型，方便数据分析师对股票异常波动进行有效的分析，为数据分析师提供了分析所需的必要数据及信息，使得数据分析师在能正确理解数据意义的基础上做出高效的分析。

(4) 本发明针对舆论数据的关键信息高维且稀疏的问题，提供了一个通用的数据分析方案，提高了数据分析师的分析效率。

**附图说明**

图1为本发明挖掘舆论中与股票异动相关的情感特征的过程示意图。

图2为爬虫采集数据的流程示意图。

图3为文本分布式表示及CNN网络的连接示意图。

图4为词向量表示的文本特征提取过程示意图。

图5为异动发生前的发帖量变化示意图。

**具体实施方式**

为了更为具体地描述本发明，下面结合附图及具体实施方式对本发明的技术方案进行详细说明。

本实施方式以提取舆论中与异动类型为“跌幅偏离值达7%的证券”相关的特征为例子，如图1所示，该提取过程包括文本的采集和标注、词语的分布式表示与情感特征提取、发帖量统计和大盘变化统计等步骤，具体实现过程如下：

（1）对舆论数据进行采集，并根据股票异动类型对样本进行类别标注，将数据集分为训练集、验证集和测试集。

根据步骤（1）对舆论和异动类型、异动发生的日期进行采集，并根据异动类型对样本进行标注。其中舆论数据和股票异动类型与异动时间通过图2所示的爬虫结构进行采集，对原始文本数据做预处理时，根据文本的句子长度的分布信息，定义涵盖98%以上的句子长度的一个句子长度值的取值范围，去除异常值或将超出范围值整理到正常范围中，预处理后得到规整的数据，可以选择不同的词嵌入方法，不同的算法进行灵活的分析。

（2）构建文本分词时所需的金融领域的词典和词语情感表示时所需的情感字典和程度副词词典。

为了充分考虑文本的情感和语义，根据步骤（2）构建所需的词典，其中证券字典如表1所示，部分情感字典如表2所示：

表1

|  |  |  |  |
| --- | --- | --- | --- |
| 成交数量 | 成交笔数 | 跌势 | 日最高价 |
| 探底 | 日交易额 | 关卡 | 空头 |
| 反转 | 多头 | 卖压沉重 | 涨势 |
| 高价区 | 骗线 | 开低 | 开平 |

表2

|  |  |  |  |
| --- | --- | --- | --- |
| 情感词 | 值 | 情感词 | 值 |
| 萎缩 | N | 企稳 | P |
| 减持 | N | 飘红 | P |
| 回落 | N | 走高 | P |
| 微跌 | N | 看好 | P |
| 空仓 | N | 扭转 | P |
| 暴涨暴跌 | N | 良机 | P |

（3）构建了情感矩阵和词嵌入两种文本结构化表示方案，把分词后的文本数据表示为向量；对样本进行分词，方案1利用word2vec方法将每个词转换成向量，或根据上述词典通过方案2将每个词转换成向量(非零即1)。

分词时导入步骤（2）构建的证券字典，采用结巴分词工具将句子分割成词语集，如图3所示，再根据步骤（3）的方案一用word2vec表示文本中的词语，即对于每个句子，根据语义把句子中的词语转化为高维的词向量，将语义相近的词语置于相近的位置；方案二，则基于情感词典和程度副词词典将句子中的词语表示为数据向量。

（4）提取数目不同的句子组成的文档的特征。

提取文档的特征时，先把数字矩阵形式表示的句子输入到RNN网络(即一个Encoder过程，由一个输入层和一个隐藏状态层组成，隐藏层的隐藏节点个数为文档的词语的数目)中，即采用RNN算法将长度不同的文档表示为固定大小的特征向量，模型的输入为维度不同的数字形式的词语向量矩阵，输出为固定大小的特征向量。

（5）提取一个日期戳对应的所有的文本文档的特征。

在提取一个日期戳对应的所有文档的特征过程中，先将每个文档对应的特征向量直接拼接在一起，输入到CNN网络(如图3所示，CNN包括一个输入层，输入层的维度为一个日期戳对应的文档数目文档特征向量的维度；一个卷积层，卷积核的数目是16，卷积核的大小为一个日期戳对应的文档数目256；一个最大池化层和一个全连接层，全连接层的节点数为128)中，生成一个日期对应的特征向量。

（6）提取异动发生之前30天的舆论文本的特征。

在提取一种异动类型对应的舆论文本内容的情感特征过程中，将异动发生之前30天的特征向量直接拼接在一起，输入到LSTM网络(输入层的timestep为30，隐藏层节点数为128，最后一个隐含层作为输出层)中，即采用LSTM对多个日期对应的整个文本的情感特征进行提取，生成一种异动类型对应的特征向量，整个过程如图4所示。

（7）构建股票异动类型分类模型，如图1所示，使用步骤（6）的特征、发帖量变化和大盘变化特征送入全连接层，再通过分类器得到异动类型分类的结果，反复迭代训练出模型。

本实施方式将异动发生之前的舆论内容的特征向量、发帖量变化和大盘变化送入全连接层，再通过SVM分类器得到异动类型的分类结果，反复迭代步骤（4）到（7）的过程训练模型。

（8）用测试集对模型的性能进行评估。

最后可以调用模型，输入舆论文本数据、发帖量变化和大盘变化，生成异动类型的分类结果，异动发生前的发帖量变化如图5所示。

上述对实施例的描述是为便于本技术领域的普通技术人员能理解和应用本发明。熟悉本领域技术的人员显然可以容易地对上述实施例做出各种修改，并把在此说明的一般原理应用到其他实施例中而不必经过创造性的劳动。因此，本发明不限于上述实施例，本领域技术人员根据本发明的揭示，对于本发明做出的改进和修改都应该在本发明的保护范围之内。

**说 明 书 摘 要**

|  |
| --- |
|  |

本发明公开了一种基于舆论数据的股票异动分析方法，其建立了一套文本情感分析的模型，通过自然语言处理技术，提取海量的舆论数据的特征，挖掘出与特定的股票异常波动类型的特征，帮助领域专家和数据分析师分析股票的行情，帮助投资者获取关键信息，改善投资策略，并给定通用的方法，大大提高金融文本数据分析的效率，符合领域专家及投资者的需要。此外，本发明对于不同的股票异动结果类型均适用，能够指导领域专家对舆论的情感特征进行提取，给出情感分析所需要的数据及信息，使得领域专家在股票异动检测的工作中尽量发挥其领域知识方面的优势，更加全面准确地掌握有效信息，有助于改善投资策略。

**说 明 书 附 图**

|  |
| --- |
|  |

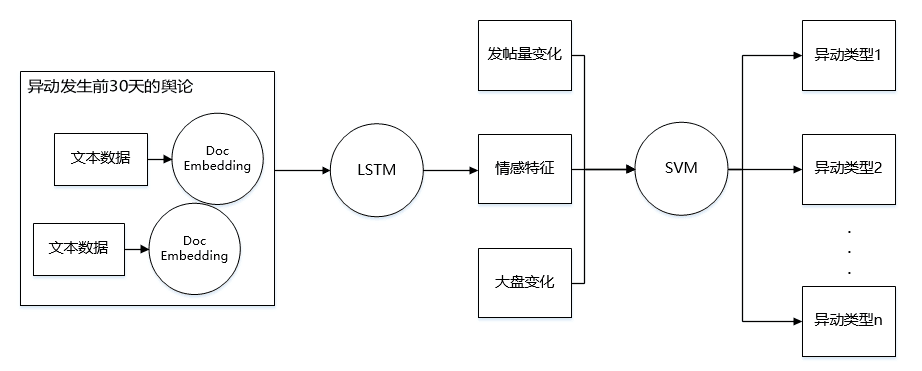


图1

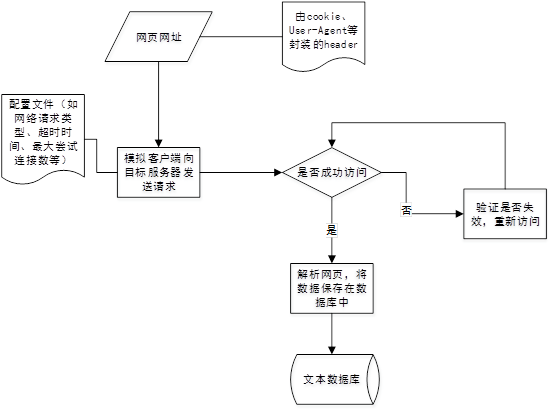


图2

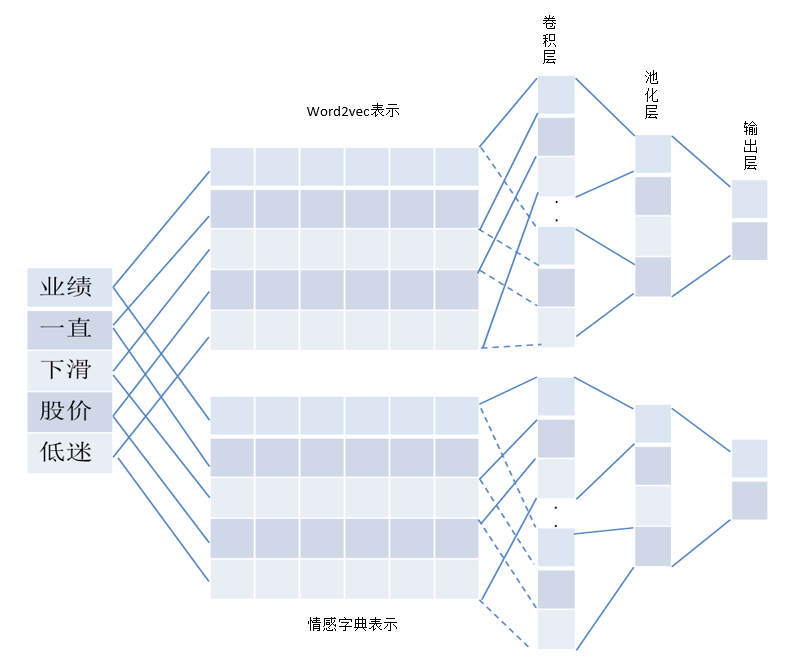


图3



图4

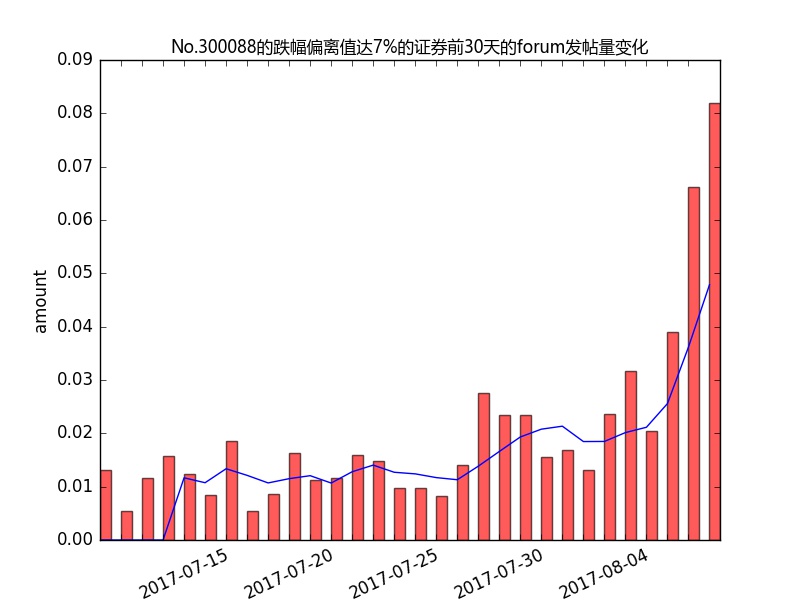


图5