1. 投资者情绪指标的构建
   1. 数据处理总流程

本文通过爬取网络股评文本、运用最新的深度学习技术对股评文本进行分类、识别股评看涨看跌情绪等步骤，并最终以适当的数学模型构建出股评投资者情绪指标。希望通过这一指标能够较为全面的、实时的反映大部分股市投资者的看涨和看跌的情绪程度。

其中网络股评文本数据全部爬取自东方财富网股吧。据Alexa[[1]](#footnote-1)的最新数据显示，东方财富网全球综合排名位于1707位，中文网站15位，金融财经类网站第1位。而其股吧是国内公认的影响力最大、讨论度最高的股市讨论社区，覆盖了国内较多的投资者人群，其中所蕴含的投资者情绪应能大体反映国内总体的投资者情绪。因此，本文选择东方财富网股吧中的帖子及回复文本作为识别情绪的源文本。

在东方财富网股吧中，以关键字分成若干个子论坛，或者称为“股吧”，子论坛的讨论内容主要是围绕关键字进行的，比如：“上证指数”、“深证成指”分别为讨论上证指数和深圳指数的子论坛，而“大话财经”则是用户讨论财经新闻的子论坛。同时，股吧中对于沪深两市中的每个个股都有以个股名字和代码为关键字的子论坛，其中主要讨论的内容是围绕这个个股进行的。因此，本文选择以“上证指数”和“深证成指”股吧中的文本构建大盘情绪指数，以沪深300所有成分股的个股吧中的讨论文本构建个股情绪指数。

本文构建情绪指标的步骤共可分为三大步，分别是网络股评文本的爬取与预处理、文本情绪分类与识别和情绪指标的构建。其总体流程如图1所示。

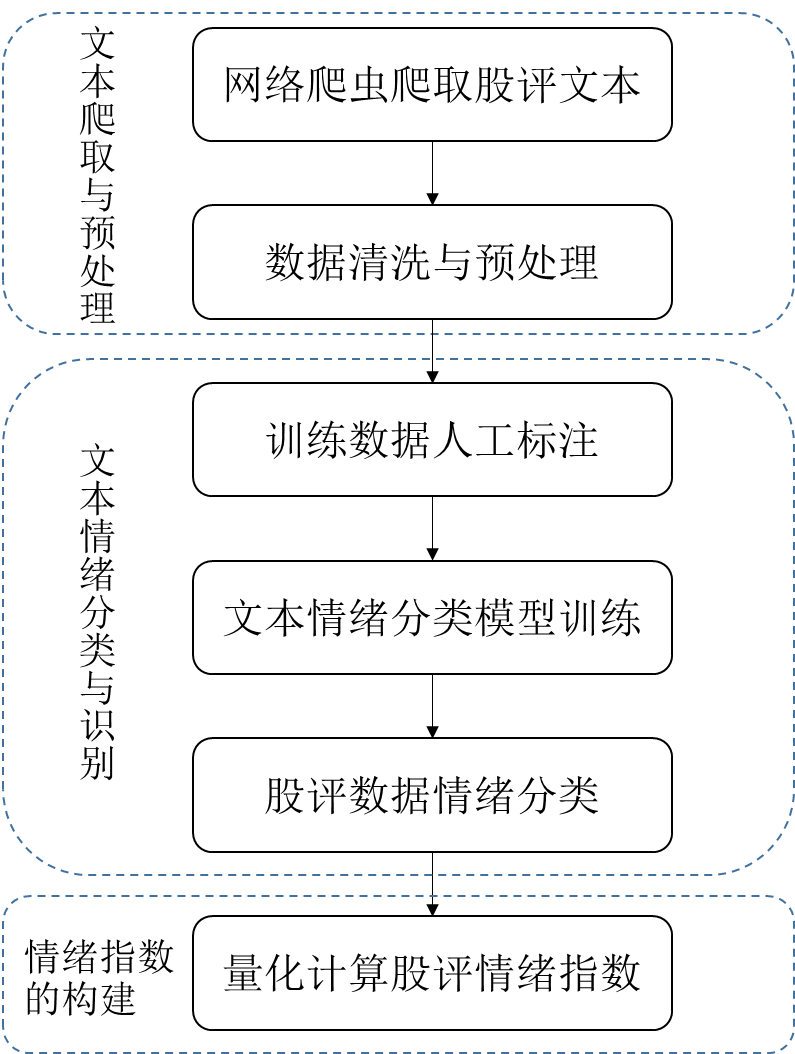


图1 数据处理总流程图

以上处理流程前两大步所涉及的所有代码限于篇幅关系未能在本文列出，但已在代码托管网站github上开源[[2]](#footnote-2)。以下分别对每个处理流程进行具体的介绍。

* 1. 股评文本的爬取与预处理

股评文本的爬取与预处理主要分为网络爬虫爬取股评文本和数据清洗与预处理两步。

* + 1. 网络爬虫爬取股评文本

网络爬虫爬取股评文本主要是通过python构建爬虫程序，自动的从东方财富网股吧中爬取相应的股评文本。由于东方财富网股吧是开放性论坛，注册用户和非注册用户都可以正常浏览所有内容，因此python程序只需通过一般的http访问请求即可抓取到所有内容。爬虫程序的主要代码流程如下：

1. 获取并构建需爬取的股票代码列表及每个个股需要的时间区间
2. 从需爬取的列表中选取一个未爬取的股票代码，构造该股票代码对应的股吧的URL链接，通过python代码获取股吧的帖子列表。
3. 从帖子列表中选取一个未爬取且帖子发布时间在所需时间区间内的帖子。
4. 利用python的BeautifulSoup库分析帖子内容，提取帖子的标题、作者、发布时间、内容、所在股吧等内容，获取帖子的回复列表。
5. 从回复列表中提取一个未解析的回复，分析内容，提取回复的作者、回复时间、内容、所回复的帖子等内容。
6. 循环迭代(5)，直到回复列表里的所有回复内容被提取。
7. 循环迭代(3)-(6)，直到所有需爬取的帖子都被爬取过。
8. 循环迭代(2)-(7)，直到所有需爬取的股票其股吧都被爬取过。

本文希望构建能够反映沪深两市总体情绪的个股情绪指数，因此选取沪深300所有成分股为所需爬取的股票，其在2015年1月1日至2018年10月1日期间作为沪深300指数的时间区间作为所需爬取的时间区间。而大盘情绪指数以“深证成指”和“上证指数”两个股吧中的文本构建，爬取区间皆为2015年1月1日至2018年10月1日。

* + 1. 数据清洗与预处理

数据清洗与预处理主要目的是为了剔除异常数据，整理数据格式，以方便进一步处理。其主要分为三步：

1. 剔除重复帖子及回复。由于爬取时是以一页页的方式爬取数据，获取上页和下页的列表有时间间隔，因此上页的内容可能会重复出现在下页中。
2. 替换文本中的所有空白字符（如换行符、制表符等）为空格，多个空白字符压缩为一个。
3. 筛选帖子内容、回复内容，只保留文本长度小于等于256的帖子和回复。

数据清洗的第(2)、(3)步主要是为了下一步对文本进行情绪分类和识别。其中第(3)步因为现有的分类模型对较长的文本难以分类准确，而且更重要的是，文本太长会使得难以训练分类模型。

经统计，本文爬取的文本经第(2)步数据处理后，共有20686493条个股股吧帖子、50199652条个股股吧回复、3184325条大盘股吧帖子、7045306条大盘股吧回复。文本长度小于等于256的有19685759条个股帖子、49807354条个股回复、3024948条大盘帖子、6989917条大盘回复，分别占比95.16%、99.22%、94.99%、99.21%。因此可以看出，本文即便通过长度过滤掉了一些数据，但仍保留了绝大多数的数据，保留的数据所作出的分析结果应能较好的反映总体数据的情况。

* 1. 文本情绪分类与识别
     1. 深度学习技术与谷歌BERT模型

深度学习是机器学习中表征学习方法的一类。深度学习技术的主要原理是：训练数据由特征和其该特征对应的标签组成，然后以这些数据通过特定的训练算法训练多层人工神经网络模型，而神经网络模型可以看成一组含参的非线性函数，通过训练参数，这一非线性函数可以“拟合”（或称学习）训练数据的特征映射到标签的关系，并可以使用拟合到的关系去预测未标注的数据标签。

图2展示了基本的三层全连接神经网络模型，箭头表示数据流向，圆圈表示一个神经元。而每一个神经元的输出可以表示成如下公式：

其中，为一个非线性函数，称为激活函数，为待定的参数，为前一层第j个神经元的输出值，若是第一层，则为第j个输入数据。

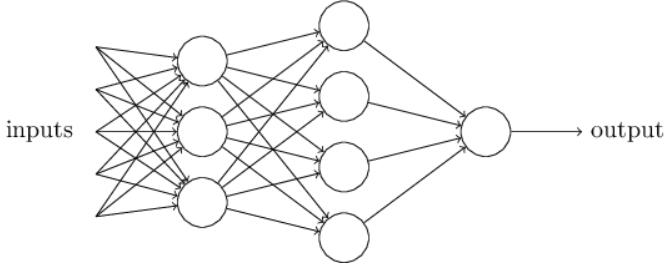


图2 基本的三层全连接神经网络模型

通过特定的训练算法（如常用的随机梯度下降）和训练数据，神经网络的权值会慢慢收敛到一个特定的值附近，以此能够拟合出输入数据到标签之间的关系。

而谷歌的Devlin(2018)最新提出的BERT模型提出了一种新式神经网络模型结构，这一模型采用双向转换器(transformer)，并结合注意力机制，从而在计算句中每个词的表示时能够结合词语所在的上下文。因此能够更加准确的用数字数据去表征一个句子的具体含义。这同时意味着一个类似含义的句子会与之有着相似的数字数据，从而能够更加容易的输出相同的标签。

此外，更重要的是，谷歌的BERT模型是对字进行编码并输入到神经网络结构中的，而其输出则直接是该句子的标签。因此，无需对句子和文本进行分词，直接将其输入到BERT模型中，就可以得到句子的标签。这直接解决了中文数据的分词困难问题。

因此，本文采用谷歌最新发布的BERT模型作为本文文本情绪分类的基本模型，以期望达到更好的分类效果、更准确的识别投资者情绪。

* + 1. 数据标注与分类模型设置

观察爬取到的东方财富网股吧中的帖子和回复文本，会发现大量与股评无关的文本数据，如“别影响午睡”、“祝圣诞快乐”等，这类数据无法标注成看跌、看平或看涨任一个类别，否则会使得合成出来的情绪指数不能正确的反映投资者的情绪。因此情绪分类器需要将数据分成四个类别（对于同意/反对分类器则分成三个类别），需要识别出无关的帖子和回复，以剔除无关贴对最后合成情绪指数的影响。

此外，观察回复的数据，可以发现回复的主要内容是表达了对主贴和其他回复的赞同或者反对，很少直接表达股评信息，因此还需要一个分类模型专门对这类数据进行分类。

基于本文对分类任务的要求，共需设置三个分类模型，分别对应三种不同的分类任务。三个分类模型的设置、输入、输出和其所对应的分类任务描述如表1所示。

表1 分类模型设置

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 模型序号 | 分类任务描述 | 分类类别 | 输入数据设置 | 输出数据 |
| 1 | 个股股吧帖子和回复的看涨看跌情绪分类 | 无关、看跌、  看平、看涨 | 帖子：标题+内容  回复：内容 |  |
| 2 | 大盘股吧帖子和回复的看涨看跌情绪分类 | 无关、看跌、  看平、看涨 | 帖子：标题+内容  回复：内容 |  |
| 3 | 所有吧中回复本文同意/反对分类 | 无关、反对、同意 | 回复的内容 |  |

\*注：表中的分别对应分为某一个类别的概率，且有，即单个模型的输出数据是一个四元组，四元组中的概率相加为1。

因此，为了训练三种分类模型， 需要三个标注好的训练数据集。而这三个数据集的构造方法为：

1. 从所有个股股吧帖子和回复中挑选共10000条数据，为了保证样本的代表性，每个个股吧中抽取的样本数量与其所含帖子/回复的数量成正比。从大盘吧帖子和回复中挑选共5000条数据。帖子和回复的数量比例依据两者在总体数据中的比例。
2. 由三位懂基本炒股术语的人标注所有数据，每个帖子只会标注一个类别，即无关、看跌、看平或者看涨。而每个回复文本会标注两个类别，第一类别标注这个回复无关、看跌、看平或看涨，第二个类别标注无关、对主贴表示反对、对主贴表示同意，且当第一个类别为非无关时，第二个类别标注为无关。
3. 每条数据都会被每个不同的人标注，最后的标注结果取三人的众数。三人结果各不同时，由三人讨论决定。
4. 标注好的个股股吧文本数据为模型1的训练数据，标注好的大盘股吧文本数据为模型2的训练数据，所有数据中的回复数据作为模型3的训练数据。

此外，标注数据时还考虑了以下几点：

1. 传达了利好消息/利空消息，标注为看涨/看跌。
2. 表达了自己或号召大家买入/卖出，标注为看涨/看跌。
3. 由于本文构造的投资者情绪指标主要是为了观察其对未来股市的影响，因此对于只描述现在情况或者基于当前情况发表个人情绪的文本标注为无关。如：“居然跌了”，“破18元喽”，“刚买就跌，气死我了”。
4. 预测价格，且不能看出是上升到预测价格还是下降到预测价格，标注为无关。这是由于分类器无法根据价格来判断是涨还是跌。
5. 在个股吧中讨论大盘的文本和在大盘吧中讨论个股的文本都标注为无关，由此，若大盘数据和个股数据分类器采用同一个分类器则会有矛盾，因此个股数据和大盘数据要分开训练和分类。
   * 1. 模型训练结果与对比
   1. 情绪指标的构建
      1. 帖子和回复情绪值的确定

对于每一个帖子和回复，以一个二元组刻画其所蕴含的投资者情绪，其中为帖子或回复情绪类别，有四种取值：无关、看跌、看平和看涨。为情绪值，反映其所蕴含的情绪明确程度。而此二元组的确定过程如下：

1. 分别使用分类模型1和分类模型2对所有个股和大盘股吧的帖子和回复进行分类，分类的结果使得每一个帖子和回复与一个四元组对应。分别表示其分为无关、看跌、看平和看涨的概率。
2. 使用分类模型3对所有回复进行分类，分类的结果使得每一个回复与一个三元组对应，分别表示其分为无关、反对、同意的概率。
3. 对于每一个帖子，即为其分类的最大概论所对应的类别，为其对应的概率值。
4. 对于每一个回复，若其四元组最大概率对应的类别为看跌、看平或看涨，则为其所对应的情绪类别，为类别对应的概率值。若为无关则检查三元组的概率值，若三元组的最大概率对应的类别为无关，则为无关，为0。若为同意，则与回复的帖子的相同；若为反对，则与回复的帖子情绪类别相反（若为看平则同为看平），与回复的帖子相同。

以上过程中，使用分类的概率值作为一个文本的情绪值的原因是：对于一个文本数据，若其表达的情绪清晰强烈，分类器会输出较高的概率值，而其表达的情绪含混不清，则会输出较低的概率值。

* + 1. 个股情绪指标的构建

1. 情绪指标的构建

借鉴并改进Antweiler and Frank（2004）的情绪指标构建公式，首先对于每支个股，其在时期绝对看跌、看平和看涨情绪定义如下：

其中，为个股在时期内其股吧所发帖子的看跌/看平/看涨情绪值，为看跌/看平/看涨的帖子数量，右下标为r的为回复的相应变量。

而个股在时期内的相对情绪指标构建如下：

相对情绪可认为是每一个回复和帖子所蕴含的平均情绪，再考虑发表的帖子数，得到最后的个股情绪指标：

对比Antweiler and Frank（2004）的构建公式，本文存在有如下三点改进：

1. 以每个帖子或回复的情绪值作为权值进行相加，原公式可认为权值都为1。但由于情绪表达清晰的文本更容易被人理解，也更容易影响到他人，因此理应对最后的总情绪值有更大的贡献。
2. 原公式没有考虑回复的情绪，本文考虑了回复所蕴含的情绪。但一般来说，帖子会比其回复更容易被其他用户浏览（如在首页可直接浏览帖子标题），因此以二八定律加权求和帖子情绪和回复情绪得到最后的绝对情绪。（可引证）
3. 参考部慧(2018)等考虑中性帖子的影响，但相比其只在加入中性帖子的数量作为乘积项，本文认为中性看平帖子能够“稀释”情绪，即在保持总体看跌/看涨的方向性不变的情况下，使情绪相对数值变小。
4. 节假日处理
5. 时期划分
   * 1. 个股情绪一致性指标的构建
     2. 大盘情绪指标与情绪一致性指标的构建
     3. 综合情绪指标的构建

* 1. 数据变量说明

表1 变量名及含义或计算公式说明

|  |  |
| --- | --- |
| 变量名 | 含义或计算公式 |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

1. 股评情绪与市场收益率关系的实证研究
   1. 样本选择与数据处理
   2. 描述性统计
   3. 平稳性检验
   4. 股评情绪对股票收益率影响
      1. 股评情绪对月度收益率的影响
      2. 股评情绪对日度收益率的影响
      3. 情绪一致性指数对股票收益率波动的影响
2. 股评情绪对市场交易量影响的实证研究
   1. 描述性统计
   2. 平稳性检验
   3. 股票情绪对市场交易量的影响
   4. 情绪一致性指数对市场交易量的影响

开盘、日内、日间收盘

盘前、盘中、前日

T,t-1,t-2

情绪一致性

1. 股评情绪对股市影响的非对称性效应研究
   1. 看涨和看跌情绪的非对称效应
   2. 极端情绪的非对称效应
2. 股评情绪在资产定价中的横截面效应
   1. 引入股评情绪的四因子模型的横截面检验
   2. 股评情绪对不同股票收益率的影响
   3. 描述性统计分析

对根据以上步骤建立完成的股评情绪指数进行整体性描述，以观察其特征。具体的描述性统计结果如表1所示。

表1 股评情绪指数的描述性统计结果

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 变量 | 样本数 | 均值 | 标准差 | 最小值 | 最大值 | 偏度 | 峰度 | Jarque-Bera统计量 |
| 盘前情绪 | 915 | 0.3382 | 0.4673 | -1.3127 | 1.6476 | -0.3473 | 2.8822 | 18.9243 |
| 盘中情绪 | 915 | -0.5793 | 0.3541 | -1.8898 | 0.6974 | -0.1712 | 3.2053 | 6.0780 |
| 前日情绪 | 915 | -0.1963 | 0.4613 | -1.8447 | 1.3782 | -0.0724 | 3.1213 | 1.359239 |
| 盘前情绪一致性 | 915 | 0.0902 | 0.0348 | 0.01953 | 0.2310 | 0.6462 | 3.4676 | 72.0156 |
| 盘中情绪一致性 | 915 | 0.0847 | 0.0302 | 0.0245 | 0.1871 | 0.2950 | 3.0068 | 13.2726 |
| 前日情绪一致性 | 915 | 0.0518 | 0.0186 | 0.0179 | 0.1866 | 1.1114 | 6.7172 | 715.1617 |

表1的统计结果显示，除前日情绪指数之外，所有变量的JB统计量都显著的拒绝正态性分布的零假设。说明各情绪样本的分布不均匀性。其中，盘前情绪样本均值大于0，说明其样本大部分大于0，结合其指标构建公式的含义，可以说明股市休盘后到股市开盘前，网络股评大多有看涨的倾向。而盘中情绪和前日情绪样本则大多低于0，说明盘中情绪和今日开盘前到昨日开盘的整体情绪大多有看跌的倾向，这一结果基本与部慧等[[3]](#endnote-1)一致。此外，三种情绪指数都存在不同程度的左偏，这说明网络股评的相对正面情绪较为集中且一般偏离均值较少，但相对负面情绪较为分散且偏离均值较多。

另一方面，从情绪一致性指数的均值可以看出，盘前的情绪较盘中情绪更为一致。而从情绪一致性指数的最大值可以看出，网络股评的情绪分歧一般比较大，较少出现比较一致的情况。

文本情绪分类综述：

文本的情绪分类（或称情感分类、情感分析等）是投资者情绪指标构建的关键一步，只有正确分类的文本才能构建出完整而正确反映投资者情绪的情绪指标。过去的研究大多采用朴素的情感词典方法或者一些经典的机器学习算法，如Zhang et al.（2011）采用的情感字典方法，Antweiler和Frank（2004）、段江娇等(2017）、部慧等(2018)采用的朴素贝叶斯分类算法，Dickinson和Hu（2015）采用的基于词向量文本表示技术和随机森林分类算法。

正如前面文献综述所述，这些经典的分类算法实现简单，但是一般对于中文数据集难以得到较为满意的结果。其主要的困难来源于中文的分词困难和词袋模型对语言的不精确建模。

而近年来发展迅猛的深度学习技术则在自然语言处理上取得了较好的效果，其主要采用的语言模型为基于词特征的词向量表示，分类模型则有循环神经网络（RNN，Recurrent Neural Network）和长短时记忆模型（LSTM，Long Short Term Memory）。但是这些方法往往只结合了单向的句子信息来理解一个词语，因此在部分复杂的数据集上也难以获得较好的效果。

2018年10月，谷歌AI团队发布了BERT（Bidirectional Encoder Representations from Transformers）模型，在机器阅读理解顶级水平测试SQuAD1.1中两个衡量指标上全面超越人类，并且还在11种不同NLP任务中创出最佳成绩，这一成果无疑是自然语言处理领域里程碑式的成就。

中文尤其是对中文数据集而言，中文与英文不同，不存在天然的分词，需要先对中文句子先进行分词才能进行建模分类，但是由于中文的特性，不同的分词往往有不同的含义，造成了句子的歧义，这使得分词算法的不正确性会影响最后分类的效果。同时，分词需要在理解句意甚至结合上下文的情况下进行，这也加大了分词的难度。卢苇和彭雅(2007)比较了朴素贝叶斯、K近邻、支持向量机等几种常见的分类算法在中文和英文数据集上的效果，结果发现由于分词的困难，中文的数据集上的准确率要大大低于英文数据集。因此，分词困难造成的分类不准确是传统分类算法遇到的普遍困难。

此外，传统的分类算法对语言的建模是一般基于词袋模型(bag of words)，这个模型认为句子的特征是单词本身决定的，而不考虑词语组成的顺序。但是这个模型对于一些特殊表意的句子则不能很好的分类，如“我喜欢你，不讨厌你”和“我讨厌你，不喜欢你”两个句子，基于词袋模型的情感字典算法和朴素贝叶斯方法可能将其分为一类，但实际上两句表达了完全不同的情感极性。这是传统分类算法的第二种困难

1. http://alexa.chinaz.com [↑](#footnote-ref-1)
2. https://github.com/FirmlyReality/guba\_sentiment [↑](#footnote-ref-2)
3. 部慧, 解峥, 李佳鸿, & 吴俊杰. (2018). 基于股评的投资者情绪对股票市场的影响. *管理科学学报,* *v.21；No.166*(04), 91-106.

   段江娇, 刘红忠, 曾剑平. 中国股票网络论坛的信息含量分析[J]. 金融研究, 2017(10):182-196.

   卢苇, 彭雅. 几种常用文本分类算法性能比较与分析[J]. 湖南大学学报(自然科学版), 2007, 34(6):67-69.

   Devlin J , Chang M W , Lee K , et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[J]. 2018. [↑](#endnote-ref-1)