利用自然语言处理技术构建中国市场金融舆情因子

江俊锋，李家豪

Abstract

本文中我们使用一种综合的算法来计算市场的舆情因子。我们运用中文语境中的自然语言处理技术，选取几大有影响力的财经网站中的新闻，对这些新闻进行情感分析，并以此构造舆情因子。我们发现，我们构造的这一舆情因子与中国市场有着显著的相关性，因此这一因子可以作为投资决策中的重要参考。

1 简介

自然语言处理是机器学习中最有前途的领域之一，而相关的研究在最近也已经取得重大的进展。许多研究也将自然语言处理的技术应用到金融市场当中。在利用自然语言进行投资决策的过程中，人们面临的主要苦难是自然语言不是结构化的数据。自然语言处理的一个主要任务就是要处理自然语言这一非结构化的数据。许多模型能够很好地将自然语言数据转换为数值数据。应用这种方法，利用自然语言数据就变得可能并且更简单了。

在这当中一种模型是基于朴素贝叶斯的。这种模型背后的逻辑是：反映出同一类情感的词总是会频繁地同时出现。这些研究往往会选择一些词作为标签词。通过研究大量文本中使用的词并着眼于标签词与其他词的出现频率之间的关系，就可以将许多不同的词进行分类。对于任意给定的文本，就能够用这些分类过的词来评价文本反映的情感。研究表明这种方法能够很好地分析推特或者新闻所反映出来的情感，而基于这种舆情因子，投资者可以相应地做出投资决策。

然而，这种方法有其局限性。最主要的局限性在于它们只关注很少量的词。一些新的词可能会表现出相似的情感，但是却由于出现频率太低而被忽视。有些时候这些词语在分析文本情感的时候会起到十分重要的作用。那么这样的信息的丢失就会对情感分析的准确性带来巨大的破坏。

这个研究主要着眼于利用更多的词进行情感分析，而非只关注少数几个典型的词。具体的研究目的包括以下几个方面：（1）自动化地从几个具有影响力的财经网站下载新闻；（2）对从网站中爬取得新闻进行预处理；（3）采用一种方法或一些算法对预处理文本数据进行分析，并最终利用每天的新闻计算当天的舆情因子；（4）选择合适的标准对舆情因子与市场趋势的相关性进行分析，评价舆情因子是否对金融投资有帮助。

文章剩下的部分安排如下。第二部分描述研究的背景和相关研究，包括结巴分词，Word2vec和WordNet。第三部分则展示研究方法与数据。第四部分包括实验结果和讨论。最后在第五部分则是我们的结论。

2 相关研究

2.1 结巴分词

中文分词比英文分词要复杂的多。进行英文分词，只需要将词语按照空格和标点符号分开即可。而中文词与词之间没有空格。因此对中文进行文本分析时，需要额外增加分词这一步骤。

结巴中文文本分词是一个中文分词模块。其算法是基于概率语言模型。它通过一个预先准备好的词典生成一棵trie树，并且计算字典中的词的词频。当处理一个需要分词的句子的时候，它会生成一个DAG（有向无环图）来记录每一种可能的分词的情况。这样一个DAG就是一个字典，字典的键是一个词起始的位置，而字典的值则是由一个词可能的结尾的位置组成的列表。

对于DAG中每种可能的分词，需要基于预先准备好的词典计算概率。然后结巴会通过从右向左的方向计算概率最大的路径。这一条概率最大的路径就给了我们一种最大可能性的分词。

对于句子中出现词典里没有的词的情况，结巴采用HMM（隐马尔科夫模型）和维特比算法来进行分词。对于不在词典中的字，它们会有四种可能的状态：B（起始字），M（中间字），E（结束字）和S（单字）。这些状态表明了这个字在词中的位置。而分词的过程主要就是基于这些状态。结巴分词的作者通过对大量文本的训练得到了三个概率表，并利用维特比算法来计算一个字最有可能的状态，并依据这种一个句子每个词的状态构成的状态链来进行分词。

2.2 Word2vec

在2013年，Google团队发表了word2vec工具。word2vec 工具主要包含两个模型：跳字模型（skip-gram）和连续词袋模型（continuous bag of words，简称CBOW）。值得一提的是，word2vec的词向量可以较好地表达不同词之间的相似和类比关系。

word2vec 自提出后被广泛应用在自然语言处理任务中。它的模型和训练方法也启发了很多后续的词嵌入模型。下面以中文为例介绍word2vec的模型。

2.2.1 跳字模型

在跳字模型中，我们用一个词来预测它在文本序列周围的词。

举个例子，“我很喜欢你呢”，这一句话可以划分为“我”，“很”，“喜欢”，“你”，“呢”。若以“喜欢”作为中心词，设窗口为2，那么在跳字模型中，我们关心的是，给定的中心词“喜欢”生成与它距离不超过两个词的每一个背景词“我”，“很”，“你”，“呢”的条件概率。

我们用数学语言来严格地描述跳字模型。

假设词典索引集D的大小为$|\mathrm{D}|$,且记D=$\{1,2,...,|D|\}$. 给定一个长度为T的文本序列，其中第t个词记为$w^{(t)}$. 当窗口大小为m时，跳字模型要求最大化任一中心词生成距离不超过m个词的背景词的总概率

$$\prod_{t=1}^{T} \prod_{-m\le j\le m, j\neq 0, 1\le t+j\le |T|} P(w^{(t+j)}|w^{(t)})$$

所以似然函数为，

$$\sum_{t=1}^{T} \sum_{-m\le j\le m, j\neq 0, 1\le t+j\le |T|} logP(w^{(t+j)}|w^{(t)})$$

最大化似然函数等价于最小化下面的损失函数，

$$-\frac{1}{T}\sum_{t=1}^{T} \sum_{-m\le j\le m, j\neq 0, 1\le t+j\le |T|} logP(w^{(t+j)}|w^{(t)})$$

我们可以用$\mathbf{v}$ 和$\mathbf{u}$分别表示中心词和背景词的向量，也就是说，对于索引为i的词，它作为中心词和背景词时的向量表示分别为$\mathbf{v}_i$ 和$\mathbf{u}_i$. 而我们要训练的模型参数就是词典中所有词的这两种向量。为了将模型参数植入损失函数，我们需要使用模型参数表达损失函数中的给定中心词生成背景词的条件概率。给定中心词，假设生成各个背景词是相互独立的，那么对于中心词w_c ，背景词w_b, b,c为这两个词在词典中的索引。那么给定中心词w_b 生成背景词w_c 的条件概率可以通过softmax函数定义为

$$P(w_b|w_c) = \frac{exp(\mathbf{u}_b^T\mathbf{v}_c)}{\sum_{i\in D} exp(\mathbf{u}_i^T\mathbf{v}_c)}$$

通过微分我们可以得到上述条件概率的梯度

$$\frac{\partial logP(w_b|w_c)}{\partial \mathbf{v}_c} = \mathbf{u}_b - \sum_{j\in D}\frac{exp(\mathbf{u}_j^T\mathbf{v}_c)}{\sum_{i\in D} exp(\mathbf{u}_i^T\mathbf{v}_c)}\mathbf{u}_j$$

上式也可以写作

$$\frac{\partial logP(w_b|w_c)}{\partial \mathbf{v}_c} = \mathbf{u}_b - \sum_{j\in D} P(w_j|w_c)\mathbf{u}_j$$

我们可以用梯度下降法或随机梯度下降法来进行迭代求解，最终求得使得损失函数最小时，词典中所有词的中心词和背景词的词向量v_i 和u_i $i = 1, 2, ..., |D|$.

当序列长度T较长时，我们可以在每次迭代时随机采样一个较短的子序列来计算有关该子序列的损失，以求得近似解。

在自然语言处理的应用中，我们会采用跳字模型的中心词向量作为每一个词的词向量。

2.2.2 连续词袋模型

连续词袋模型与跳字模型类似，它是用中心词在文本序列中前后的背景词来预测该中心词。举个例子，“我很喜欢你呢”，这一句话可以划分为“我”，“很”，“喜欢”，“你”，“呢”。若仍以“喜欢”作为中心词，设窗口为2，那么连续词袋模型关心的是，给定与中心词距离不超过两个词的背景词“我”，“很”，“你”，“呢”这四个词生成中心词“喜欢”的条件概率。

假设词典索引集D的大小为$|\mathrm{D}|$,且记D=$\{1,2,...,|D|\}$. 给定一个长度为T的文本序列，其中第t个词记为$w^{(t)}$. 当窗口大小为m时,连续词袋模型要求最大化由距离不超过m个词的背景词生成中心词的总概率.

$$\prod_{t=1}^{T} P(w^{(t)}|w^{(t-m)},...,w^{(t-1)},w^{(t+1)},...,w^{(t+m)})$$

其中，m为窗口大小，且要确保$(t-m+j)\in[1,|\mathrm{T}|]$,$j\in[0,2\mathrm{m}]$.

所以似然函数为，

$$\sum_{t=1}^{T} logP(w^{(t)}|w^{(t-m)},...,w^{(t-1)},w^{(t+1)},...,w^{(t+m)})$$

最大化似然函数等价于最小化下面的损失函数，

$$-\sum_{t=1}^{T} logP(w^{(t)}|w^{(t-m)},...,w^{(t-1)},w^{(t+1)},...,w^{(t+m)})$$

我们仍用跳字模型中表示中心词和背景词的记号，此时，对于中心词w_c以及它所对应的背景词$w_{b0},w_{b1},...,w_{b2m}$，那么给定背景词$w_{b1},w_{b2},...,w_{b2m}$生成中心词$w_c$生成的条件概率可以通过softmax函数定义为,

$$P(w_c|w_{b0},w_{b1},...,w_{b2m}) = \frac{exp(\frac{\mathbf{v}_c^T(\mathbf{u}_{b0}+\mathbf{u}_{b1}+...+\mathbf{u}_{b2m})}{2m})}{\sum_{i\in D} exp(\frac{\mathbf{v}_i^T(\mathbf{u}_{b0}+\mathbf{u}_{b1}+...+\mathbf{u}_{b2m})}{2m})}$$

通过微分我们可以得到上述条件概率的梯度,

$$\frac{\partial logP(w_c|w_{b0},w_{b1},...,w_{b2m})}{\partial \mathbf{u}_{bi}} = \frac{1}{2m}(\mathbf{v}_c - \sum_{j\in D} \frac{exp(\frac{\mathbf{v}_c^T(\mathbf{u}_{b0}+\mathbf{u}_{b1}+...+\mathbf{u}_{b2m})}{2m})}{\sum_{i\in D} exp(\frac{\mathbf{v}_i^T(\mathbf{u}_{b0}+\mathbf{u}_{b1}+...+\mathbf{u}_{b2m})}{2m})}\mathbf{v}_j)$$

上式也可以写作,

$$\frac{\partial logP(w_c|w_{b0},w_{b1},...,w_{b2m})}{\partial \mathbf{u}_{bi}} = \frac{1}{2m}(\mathbf{v}_c - \sum_{j\in D} P(w_c|w_{b0},w_{b1},...,w_{b2m})\mathbf{v}_j)$$

与跳字模型一样，我们也可以用梯度下降法或随机梯度下降法来进行迭代求解，最终求得使得损失函数最小时，词典中所有词的中心词和背景词的词向量v_i 和u_i (i=1,2,...,$|\mathrm{D}|$).

当序列长度T较长时，我们可以在每次迭代时随机采样一个较短的子序列来计算有关该子序列的损失，以求得近似解。

在自然语言处理的应用中，我们会采用连续词袋模型的背景词向量作为每一个词的词向量。

2.3 WordNet

WordNet是一个大型英文词典数据库。在WordNet中，词汇之间的关系是根据语义来相互联系的。在WordNet中词与词之间的关系主要是近义词。通过利用一个网络来表现词之间的关系，WordNet帮助我们找到词的近义词，并表现了两个词在语义的意义上有多相似。

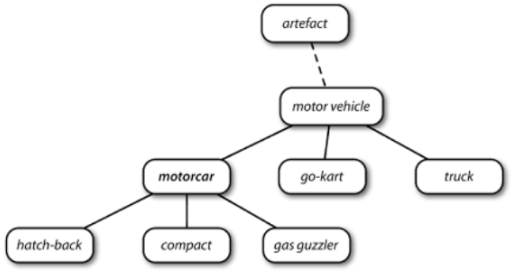

Figure 1: WordNet结构示例

由于WordNet以上的一些性质，它在进行文本的情感分析时会更可靠。汉语开放词网（COW）是由WordNet启发的大规模的、能够免费获取的汉语语义词典。它有着和WordNet相同的结构，并且是基于中文的。它包含42315个近义词集，79812个词义以及61532个词语，并且仍然在建设当中。我们的研究主要应用COW来计算每个词的情感得分。

3 研究方法

3.1 预处理

汉语，作为一种孤立语，它不是通过词形变化来表达语法，而是通过独立的虚词和固定的词序来表达语法意义，与英语相比有其特殊性，比如英语天生就具有自动分词的特性，而汉语，在一句话中并没有像英语里的空格那样划分词的符号，所以在进行中文自然语言处理之前，我们需要对其进行分词，将一篇文章切分为词袋。这是英文自然语言处理与中文自然语言处理间的一个重要的差别。对任意给定的新闻或文本，我们采用结巴分词来进行预处理。除了需要将整篇文章转换为词袋以外，实际上，还会出现许多类似“的”、“是”、“可能”……这样的助词、副词等不具有具体含义或与情感分析无关的词，这些词和符号称为停用词，它们通常会频繁地在许多新闻或文本中出现然而却没有太多的意义。为了提高精度和减少运算，在进行预处理时，我们会对其进行去停用词操作，将注意力放在对文本的情感影响大的词上面去。

要注意的是，载入语料库训练模型时，不能对语料库进行去停用词操作，因为停用词的丢失会影响正确的词向量生成，在下面的文本分析中会做具体分析。

另外，我们还可以自定义一些应该去掉的字符来减少计算量，比如去掉英文单词、数字、标点符号等。

3.2 文本分析

3.2.1 语法相似度

首先，我们要训练word2vec模型以获得足够多的词的词向量（理论上，只要用于训练的语料库足够大，我们是可以获得所有词的词向量，但实际上我们不会有这样的语料库）。为了保证一般性，我们采用中文维基百科语料库进行训练。

要注意的是我们不能对用于训练的语料库进行去停用词操作，因为停用词的存在对描述中心词也会起到很关键的作用。

为了尽可能地使每个词的词向量在其所在的超平面上线性可分，我们要设置一个合适的词向量长度。另外我们可以设置一个恰当的词频阈值，使得可以忽略一些生僻词，在不影响我们关注的词向量计算的前提下以节省内存。

训练结束后我们可以得到一个word2vec模型，与此同时，我们也就得到了所有词对应的词向量——我们只要像查字典那样在模型中查找对应的词向量即可。

现在，我们可以计算任意两个词之间的语法相似度了。比如，考虑两个词w_1 和w_2，标准化后为$w_1^{'}$ and $w_2^{'}$，于是w_1 和w_2的语法相似度为，

$$distance = \frac{w_1 \cdot w_2}{\|w_1\|\|w_2\|} = w_1^{'} \cdot w_2^{'}$$

注意：虽然我们现在可以用word2vec模型估计两个词的相似度，但是我们得到的只是它们的语法相似度。也就是说，我们只能知道某个词与哪些词语法上类似，但并不知道它们的实际语义。我们可以举一个简单的例子，考虑两个词“上涨”“下跌”，我们也经常在新闻中能看到这样的句子——“今天上证指数上涨5%”。然后你就会发现，我们可以很轻易地把“上涨”改为“下跌”而不会有任何的问题。根据**2.2**说的，这两个词的词向量基本上是一样的，这就会影响到我们去判定它们各自的情感得分。但是，word2vec的确可以帮助我们找到与给定的词在某种意义上比较接近的一些词。

3.2.2 语义相似度

在2.3中我们提到，WordNet通过一种树状结构来储存词语，这种树状结构就天然地形成一种

距离。在利用WordNet计算两个词的语义相似度时，就是利用这两个词在WordNet中的节点之间的最短路径来进行计算的。我们取两个词的节点之间的最短路径的长度的倒数作为语义相似度。每个词和自身的相似度是1，而如果两个词完全没有路径将它们的节点相连，那么这个两个词的语义相似度定义为0。这样定义得到的语义相似度是一个介于0到1之间的值，这个值越大，表明两个词之间在语义上更相似。

3.3 情感得分计算

3.3.1 标签词

我们利用一个标签词集来评估一个词的情感。其思想在于一个词与另一个我们设为标准的词的相似性可以反映它的情感。我们选择了100个在财经新闻中经常出现、并且明显反映了对待市场的情绪的词语作为标签词。为了使计算更加公平，我们选择了50个积极词和50个消极词。在计算情感得分之前，我们首先对每个词计算了它和标签词集中每个词的语法相似度和语义相似度。

3.3.2 词语情感得分计算

通过上面的计算，对于每一个词语，我们都有办法得到一个两百维的向量，这个向量的前100维是这个词语与标签词集中的100个词的Word2vec相似度，反映的是这个词和每个标签词集中的词的用法相似度，而向量的后100维则是这个词语与标签词集中的100个词的WordNet相似度，反映的是这个词和每个标签词集中的词之间的语义相似度。由于标签词集中的词就是由反映对市场的态度的词语组成，因而根据与这些词的相似度，就能够代表每一个词对市场的态度，也即情感。

在我们的研究中，我们通过每个词的这一由相似度构成的向量来计算这个词的情感得分。具体而言包括以下几个步骤：（1）先后利用Word2vec相似度与WordNet相似度，通过协同过滤的方法判断词语属于积极词还是消极词；（2）利用Word2vec相似度计算词语情感得分。具体操作过程与原因叙述如下。

对于一个词，它和标签词集里的哪些词更像，有两个维度上的含义。第一个是在用法上，也即语法层面上的相似，而第二个是语义上的相似。采用协同过滤的方法，可以找到与这个词语在语义和用法上最相似的几个词。首先根据前100维上的Word2vec相似度，找出与目标词在用法上最相似的9个词，这个过程其实就是从前100维的相似度中找出绝对值最大的9个。这样的9个词就是在用法上与目标词最接近的9个词。然后在后100维忠找出目标词与这9个词的WordNet相似度，并取出WordNet相似度最大的前3个词。那么这3个词就是在用法和语义两个层面都与目标词最相似的词。

有了这3个词我们本可以直接判断词语属于积极词还是消极词，然而仅仅判断是积极词还是消极词是不够的，事实上，举例而言，即使同样是积极词，不同的词反映的积极程度应该不同。为此，对于每个目标词，我们需要计算它们的情感得分值。计算方法是利用这3个词的Word2vec相似度的绝对值，按照这3个词是积极词还是消极词取正负，最后相加得到的值作为情感得分。

在实际操作中如前文所描述的那样，相比通过训练得到的Word2vec模型中的词语数目，COW中的词语数目仍然不够充足，因此有些时候会出现一些词语与标签词集的WordNet相似度无法计算的情况，对于这些词，我们可以认为暂时无法判断其与标签词的语义相似度，因而仅仅对用法相似度进行判断。所以对这些词的协同过滤的过程，第二步就不能在9个词中选出WordNet相似度最高的3个词，而只能取9个词中Word2vec相似度最高的三个词了。最终采用Word2vec相似度进行情感得分值的计算，也是基于相同的理由。

在这样的计算方法下，我们就可以计算出词语的情感得分。

3.3.3 文章情感得分计算与舆情因子

利用上述计算得到的情感得分，我们就可以计算出每篇新闻的情感得分，并根据一天内所有网站所有新闻的情感得分，就能够构造市场当天的舆情因子。这一过程的具体描述在这一部分阐述。

对于经过预处理的新闻，我们得到的是一个词袋。固然我们可以对词袋中的每个词计算200维的

相似度向量，并依据这个向量进行这个词的情感得分计算。然而这样会有大量的重复计算，事实上预先计算好足够多的词语的情感得分，往往就已经能够涵盖任意一篇文章的绝大多数词语，能够给我们对文章的情感作出足够准确的判断了。

为此，我们首先利用了商务印书馆出版社出版的《现代汉语常用词表》的50000词，然后我们对经过预处理的任意3000篇新闻中的词进行判断，提取出不属于《现代汉语常用词表》中的词，最终我们得到了约100000个常用词。我们预先计算好这100000词的情感得分，这样每次进行新闻的情感得分计算时，对每个词的得分就不需要进行计算的过程，而仅仅只要进行检索即可。这样的操作可能会丢失一些词语，然而结果表明对于大多数文章而言，不在常用词集中的词只占词量的5%左右，丢弃这5%的词对文章情感得分的影响并不大。

那么，每篇文章就利用这一常用词集中的得分，加总文章中词的情感得分，并除以词数，从而得到整篇文章的情感得分。对于一天中所有文章的情感得分，加总后除以当天统计的文章数，就得到当天的舆情因子。这一舆情因子就能够很好地反映市场的情绪了。

3.4 相关性分析

我们需要建立一个标准来评价舆情因子是否能够很好的反映市场情绪。我们采用的主要方法是利用舆情因子与市场的一些指标进行线性回归。假设市场情绪与市场走势是正相关的，那么舆情因子与市场指标做回归的结果的显著性就能反映舆情因子效果的好坏。

4 实验结果和讨论

我们计算了181天的舆情得分，并用这个舆情得分与一些大盘指数比如上证指数和深证成指进行相关性分析。

对它们进行线性回归，我们可以看到截距项以及一次项系数都很好地通过了显著性检验。

另外求得皮尔逊相关系数为xxx，说明我们的舆情得分与大盘指数呈中等程度的相关性。

5 结论

本文中我们建立了一种算法来计算中国市场的舆情因子，这一舆情因子与中国市场有着显著的相关性。这个因子能够为我们作投资决策提供一种新的方法。

本文的主要贡献在于建立了计算市场舆情因子的方法。如果我们能够计算各种金融产品的舆情因子，那将会更加有帮助。同时，将舆情因子与传统金融因子相结合能够帮助我们作出更好的投资决策。我们非常期待能看到相关的研究。

6 参考文献

[1]George A. Miller (1995). WordNet: A Lexical Database for English. Communications of the ACM Vol. 38, No. 11: 39-41.

[2]Christiane Fellbaum (1998, ed.) WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

[3]Building the Chinese Wordnet (COW): Starting from Core Synsets. In Proceedings of the 11th Workshop on Asian Language Resources: ALR-2013 a Workshop of The 6th International Joint Conference on Natural Language Processing (IJCNLP-6). Nagoya. pp.10-18.

[4]Theoretical and Practical Issues in Creating Chinese Open Wordnet (COW). Paper presented at The 7th International Conference on Contemporary Chinese Grammar (ICCCG-7), Nanyang Technological University, Singapore.

[5]Linking and extending an open multilingual wordnet. In 51st Annual Meeting of the Association for Computational Linguistics: ACL-2013. Sofia. 1352C1362

[6]Rao T, Srivastava S. Analyzing Stock Market Movements Using Twitter Sentiment Analysis[C]// International Conference on Advances in Social Networks Analysis and Mining. IEEE Computer Society, 2012:119-123.

[7]Kim Y. Convolutional Neural Networks for Sentence Classification[J]. Eprint Arxiv, 2014.

[8]Day M Y, Lee C C. Deep learning for financial sentiment analysis on finance news providers[C]// Ieee/acm International Conference on Advances in Social Networks Analysis and Mining. IEEE, 2016:1127-1134.

[9]Zhang W, Skiena S. Trading Strategies to Exploit Blog and News Sentiment[C]// International Conference on Weblogs and Social Media, Icwsm 2010, Washington, Dc, Usa, May. DBLP, 2010.s

[10] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119).

[11] Alec Go, Richa Bhayani, Lei Huang, 2009 [R]. Twitter Sentiment Classification Using Distant Supervision.

7 附录

7.1 图表

Correlation between Senti-score and market indexes

Time series of Senti-score and market indexes

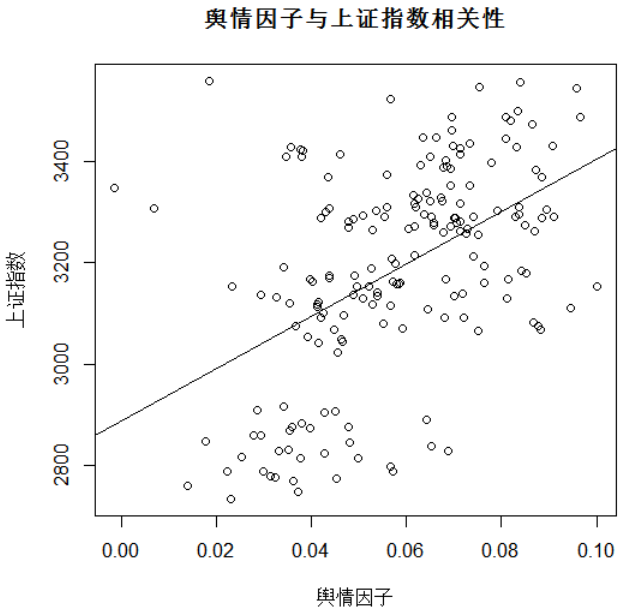

Figure 2: 舆情因子与上证指数相关性图像

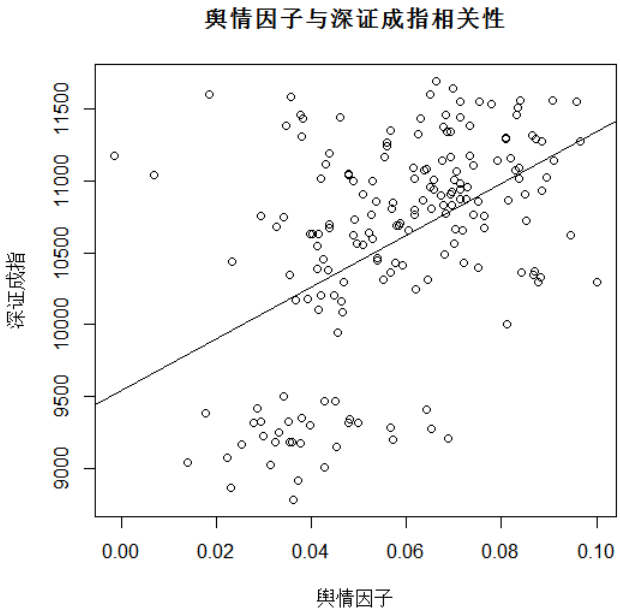

Figure 3: 舆情因子与深证成指相关性图像

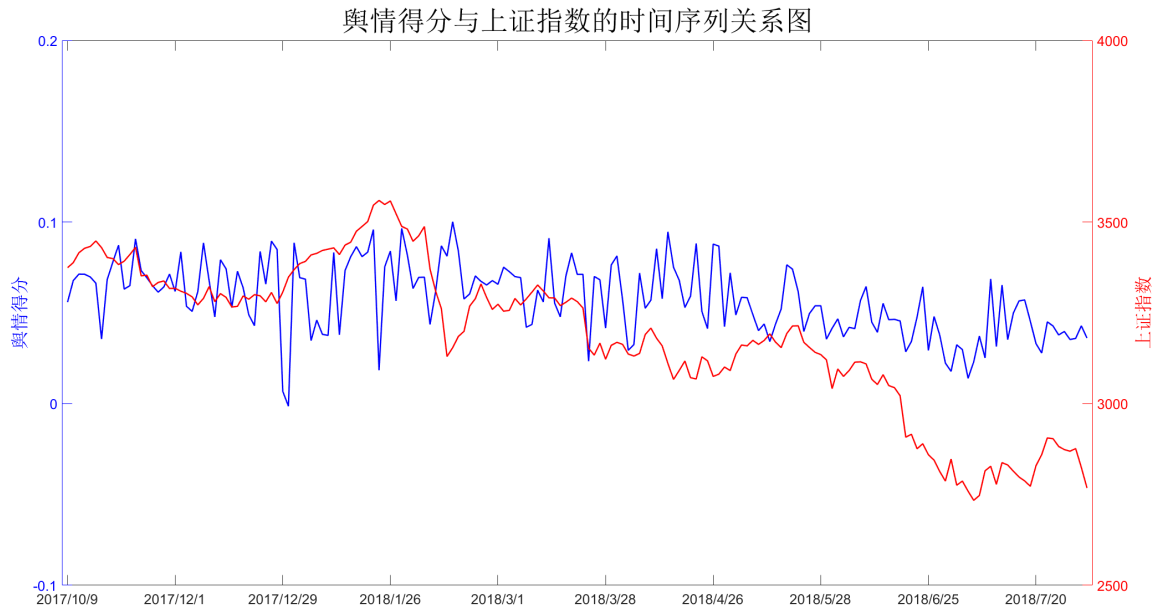

Figure 4: 舆情因子与上证指数的时间序列

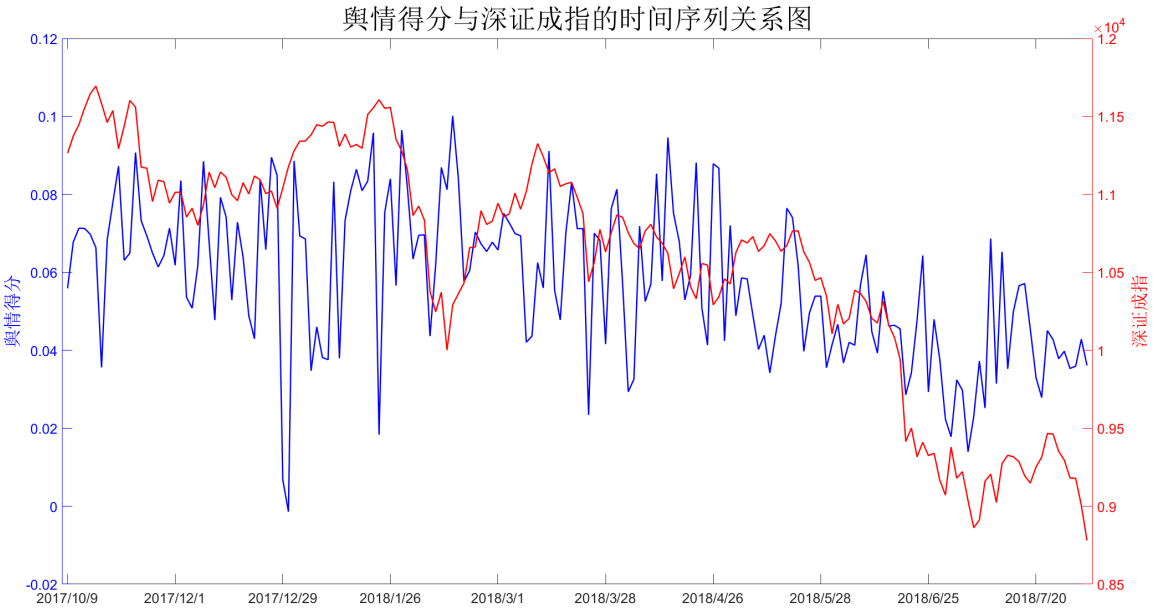

Figure 5: 舆情因子与深成指数的时间序列