



审稿意见汇总

2958-1478/© Shuangqing Academic Publishing House Limited All rights reserved.
Article history: Received August 28, 2023 Accepted September 14, 2023 Available online September 15, 2023
To cite this paper: 李金阳, 金明哲, 宿久洋(2023). 基于预训练语言模型的股市变动预测的分析. 人工智能研究, 第 1 卷, 第 2 期, 26-39. Doi: <https://doi.org/10.55375/aif.2023.2.3>

	详细
请您简要描述论文的研究内容和意义, 并评价, 特别是, 文章提出了那几个重要的观点或方法、问题是否成功解决、对本学科有什么贡献?	本文研究如何利用预训练语言模型来预测股市变动。本文尝试解决股市预测任务中, 新闻语料特征少噪声多、超长文本处理等挑战。文章的创新点包括提出了一种文本提取算法来处理超长文本, 以及对不同处理模式进行对比实验。通过实验结果, 作者发现, 所提出的文本平均提取算法能够提高预训练语言模型约 3%的准确度, 在股市预测任务中具有优势。
您觉得文章的优点、缺点有哪些? 主要理由是什么? 论文结果不正确或有重大失误是什么? 论文缺乏重要的结果是什么? (需重点关注模型和推导过程的准确性)。	问题和建议: 1. 第一部分最后提到搭建了几种模型, 需要补充具体是哪种模型。 作者修订: 基于词频特征, 基于静态词嵌入和基于动态词嵌入的几种常见文本处理模式。 2. “为解决多义词和长文本遗忘的问题, 2018 年 Devlin 等提出了 BERT 模型”-需要引用这个文献 作者修订: 已补充, 见文献[5] 3. 关于文本提取算法 这个计算公式是否意味着涨跌日的样本(交易日)应该数量相同呢? 不然可能存在偏差 比如, 样本中, 只有 1 天是上涨的, 1000 天是下跌的, 那么, 得到的结果就会出现偏差 在实验中是如何处理这个问题的呢? 作者反馈: 设置了 r 这个值来消除样本差的影响。 举个例子, 比如一个词在上涨日语料中出现 5 次(np)下跌日出现 1 次(ni), 上涨日语料总共有 500 词(n)下跌日语料总共有 100 词(i)。(方便理解先不考虑 k) 如果不设置 r 值则 $ps = np - ni / (np + ni) = 4/6$ 但是从词频来看其实这个词在二者当中出现频率差不多并不应该拥有这么高的 ps 所以设置 r 上涨日新闻的总词数和股价下跌日新闻的总词数的比 = $n/i = 500/100 = 5$



则 $ps = |np - r*ni| / (np+r*ni) = 0/10 = 0$

等于 0 就正好可以表示这个词在两者当中出现频率相同。

在实际情况中，为了避免个别词仅在上涨日或者下跌日出现几次的极端情况，会设置 k 值来降低这些极端情况词的影响，例如 $k=5$ 时（具体实验使用的 5，其他看语料大小还有想要的效果设置 10, 100 应该都行），那些出现仅一两次的词的 ps 就能够有效减小，而对那些词频比较大达到几千上万的词几乎没有影响。

另外总体来说，上涨日和下跌日的比例是接近 1:1 的。

4. Volatility Score (VS) 部分，提到的 PS 应该都是 VS？ 请作者核对是否书写有误

作者修订：已修改为 VS

5. (13) 公式括号里面 “+1” 是什么意思？

作者回复：这里用 1 做基准，因为对于 $y=x^2$ 来说， $x>1$ 的时候 y 大于 x ， $x<1$ 的时候 y 小于 x 。

设置这个的目的是希望能够把股价变动率比平均大的日子的 vs 也变大。尤其是那些达到 2%，3% 甚至 5% 的日子出现的词。

整体而言所有样本的股价平均变动率在 0.7% 左右，取 0.7。大部分的股价变动率在 0.7 上下，能够达到 2% 的变动率的日子很少。而恰恰这种日子中出现的词很重要。

所有样本的股价平均变动率在 0.7% 左右，取 0.7。

而对于那些变动小于 0.7% 的（例如某天变动了 0.1% 取 0.1）， $(\Delta p_j - \Delta p + 1)^2$ 就会接近于 0（例如 $(0.1 - 0.7 + 1)^2 = 0.16$ ）。

而对于那些变动约等于平均的日子出现的词被认为提供的信息不重要，则 $(\Delta p_j - \Delta p + 1)^2$ 就会接近于 1（例如 $(0.7 - 0.7 + 1)^2 = 1$ ）。

而对于那些变动约大于平均的日子出现的词被认为提供的信息很重要（例如某天变动了 3.0% 取 3.0），变动越大越被认为重要，则 $(\Delta p_j - \Delta p + 1)^2$ 就会大于 1（例如 $(3.0 - 0.7 + 1)^2 = 10.89$ ）。

最终计算词在整个语料中的 vs ，就可以根据 vs 的大小来判断这个词对于变动是否重要。

6. Δp 是什么指标？需要解释

作者回复：已补充。平均的定义是所有日期的变动率的绝对值的和除以日期数。

7. 关于涨跌问题：

股票的涨跌这里是按日，涨的概念是：当日收盘大于昨日收盘。PS\VS 也



是的计算也是基于这个逻辑。

而下方股价的变动是指，当日的收盘价格大于开盘价格

如上两个逻辑不一样，存在样本和度量的指标不是一回事的问题。

这意味着，PS/Vs 是用于描述：当日收盘价大于前日收盘价 这个关系的特征，而下方要研究的是当日收盘价大于当日开盘价的关系，即，特征值和目标存在差异，请作者对这个问题进行补充和解释

作者修订：

所有的标注都是基于股价的变动，这里不应该写按日的涨跌。

即股价变动的标记和 ps/vs 的计算都是基于对应日的股价变动（对应日收盘价减去开盘价）。这里的变动幅度也是基于此计算的。

语言不够严谨，很抱歉造成了歧义。

本研究的目的是对于股票变动的预测 (Stock market movement prediction, 株価変動予測)。

“股票的涨跌这里是按日，涨的概念是：当日收盘大于昨日收盘。”

这个概念在现有阶段均未使用。

之前所有提到的均为误用。

8. 关于“使用网络爬虫在路透社官网爬取，总计 189,613 篇”爬取的也是英文金融新闻吗？爬取的新闻和公开的数据是否有去重？

作者回复：同一个网站同一个版面的英文金融新闻，爬取的部分（2020 年 9 月-2020 年 12 月）和公开的部分（2009 年 9 月-2020 年 8 月）不在同一时间段，不包含重复内容。

9. 准确率

关于准确率（4.6）这里，请作者考虑如下问题：

1. 预测 t 日涨跌，从实用角度，应该是基于 $t-1$ 日数据（新闻语料）

2. 如果采用 t 日数据（新闻语料）来和 t 日的涨跌，是否缺乏科学性？

比如，2023.3.2 日，股市大跌，当日大量的语料都是报道跌的消息，因此，类似，惨、亏等词肯定在当天的语料中大料提及。但是，并不是因为提及这些词导致跌，而是，因为跌才出现这些词汇

因此，从预测股市变化角度，应该使用的是 2023.3.1 日的语料来预测 2023.3.2 日的涨跌

当前文章中针对这个语料时间的说明不详，请斟酌修订

作者回复：日期所对应的语料具体指的是前一日交易日开盘后到当前交易日开盘前的所有新闻。



	举个例子，例如在 9 月 7 日是夏季美东时间 (UTC -4) 9 点 30 开盘，则在前一交易日 9 月 6 日 9:30 (UTC-4) 到 9 月 7 日 9:30 (UTC-4) 之间发布的新闻都会被用来预测 9 月 7 日的变动。 在冬天例如 2023 年 1 月 2 日新年休市，则 1 月 3 日的动向由 1 月 1 日冬季时间 (UTC-5) 9:30 到 1 月 3 日的 (UTC-5) 9:30 之间发布的新闻来预测。 具体标注使用了 <code>python</code> 下一个叫 <code>Arrow</code> 的包来实现的。 也就是说实际上不是固定的哪一天的语料。但是也不会出现预测时间段内的新闻用来预测。
您认为文章是否存在写作上的问题比如：错别字多、语法问题大、语言不够顺畅等。	部分语句可以优化，见标注 作者：接受调整建议
若有其他问题，请在这里补充	

作者反馈：
已调整，调整思路见表格。

人工智能研究
Artificial Intelligence Frontier
 ISSN: 2958-1478 (print)
 雙清學術出版社
 Contents lists available at www.qingpress.com
 Journal homepage: qingpress.com/zh-cn/journals/45