雙清學術预印本
PrePrints

雙清學術出版社
Contents lists available at www.qingpress.com
Journal homepage: qingpress.com/zh-cn/journals/26

基于 K-BERT-BILSTM 的股指预测的研究

张源龙 [1] 吴梓杰 [2*]

1:张源龙,齐鲁工业大学,研究方向:计算科学， Email:3269388464@qq. com
2:吴梓杰,齐鲁工业大学,研究方向:金融保险， Email:florian492@outlook. com
*通讯作者：吴梓杰

摘要：中国证券市场由于历史遗留原因存在着“政策市”的诟病。相比较于西方发达经济体，我国金融证券市场在资产价格定价上更多地受到政府宏观经济调控的货币政策和财政政策的影响。亦有有学者认为导致该现象的一部分原因是不成熟理性的投资者行为。故在此，我们基于文本大数据和自然语言处理技术改进了 K-BERT 模型,并将其应用与对中国股市的投资者情绪和政策不确定性衡量当中；并基于该指标建立了双向 LSMT 神经网络模型对上证指数的走势进行了预测，将二者结合在一起的模型具有更高的准确性。

关键词：事件研究,政策不确定性,自然语言处理

我们基于 Weijie Liu (2020); 许雪晨 and 田侃 (2021) 的研究成果[1,5]，提出了 SA-LSTM-K-BERT 模型，并将其应用于中国金融市场的投资者情绪和政策不确定性衡量当中。本研究在 Weijie Liu (2020) 的基础上改进了原有模型，使其能够进行情绪分析的任务。在许雪晨与田侃 (2021) 的研究结果上改进了其 SA-LSTM-BERT 模型，为其增添了包含了中国证券市场先验知识的知识层（Knowledge Layer)，增强了其在金融领域的文本情感分析能力。

后文行文结构如下：第二部分为文献回顾，第三部分主要介绍了 BERT 模型及本研究中所改进的 K-BERT 模型，第四部分介绍了实验数据的来源，第五部分为实验结果的实证回测，最后一部分为文章结论。

1 文献回顾

1.1 股价行为研究

投资者情绪对公司股价的影响是行为金融与资产定价的一个重点领域，杨春鹏 (2007) 认为：作为直接做出决策的主体，投资者情绪对股价的走势在很大程度上有着一定的影响[8]。随着行为金融学的理论的发展和实证研究，有学者认为 Ritter (2003) 投资者的非理性行为也可能影响股票价格的波动[13]。Liu and Zhang (2015) 检验了政策不确定性指数对股价指数的可预测性，结果显示将政策不确定性指数纳入解释变量有助于

Article history: Available online October 16, 2022
To cite this paper: 张源龙,吴梓杰(2022).基于 K-BERT-BILSTM 的股指预测的研究. 雙清學術预印本，卷 1，第 1 期, 9-15.
Doi: https://doi.org/10.55375/preprints.2022.1.2

提高预测模型的精确度。这使得把投资者情绪纳入股价的解释变量，并基于此在投资中获得更高的收益率成为了可能[14]。

中国证券市场由于历史问题的原因有着“政策市”的诟病。从国内学者在该方面的实证研究来看，王琳 et al. (2021) 在一项最近的研究中运用 EGRAH 模型对央行货币沟通政策对金融资产价格的影响进行了分析，结果显示央行沟通对股票市场的价格影响是显著且合意的；这说明了宽松的货币政策会导致股价市场的上升，反之，收缩的货币政策可能导致股市的下跌[11]周学伟 et al. (2020)基于多因子混频波动率模型，发现货币政策不确定性会显著增强证券市场中的行业波动[7] 。

黄虹 et al. (2021)在研究中发现，在经济下行阶段，政策不确定性减少了企业投资。据此，黄虹认为投资者情绪是政策不确定性影响企业投资的中间指标[10]。此外，杨春鹏 (2007) 在对政策效应根源的研究中，从投资者的过度自信等非理性行为的影响的角度解释了我国证券市场上的异常现象。杨春鹏认为：不成熟的投资者是导致我国证券市场政策效应显著的原因。这从另一个方面反映了想要全面的考察中国证券市场的行为，仅考虑政府和投资者单独一方是不完整的，若将二者的影响效应结合起来，可能会获得更显著的结果[8]。

1.2 自然语言处理

本研究主要应用了深度学习中的自然语言处理技术中的情感分析方法。自然语言处理技术是现代人工智能技术研究的一个主要范畴。自然语言处理技术通过机器学习来处理和理解文本和数据，进而根据目的进行不同的处理。

Pang et al. （2002）首次将情感分析应用于时间序列预测上。Pang 的分析方法主要是基于语义规则的情感分析。这种分析方法的优点在于分析方法建模简便。但困难之处在于该方法主要是将分词后的词语与已拥有的情感词典进行比对评估，而情感词典的搜集十分困难；除此以外这种分析方法过分强调对于词典的依赖性，从而忽略了文本整体下的语境意义[9]。姜富伟 et al. （2021）在基于前人研究的基础上通过人工筛选和 word2vec 算法构建了一个较为全面的中文金融情感词典。他们发现，媒体文本的情绪在更准确地衡量我国投资者情绪的同时，提升了模型的样本外与样本内预测能力。由于金融词汇具有专业性的特点，这使得一般的情绪词典可能无法很好地衡量投资者的情绪指数，故姜富伟等人的研究弥补了我国缺乏一个公开且被广泛接受的中文金融情绪词典的空白[2]。

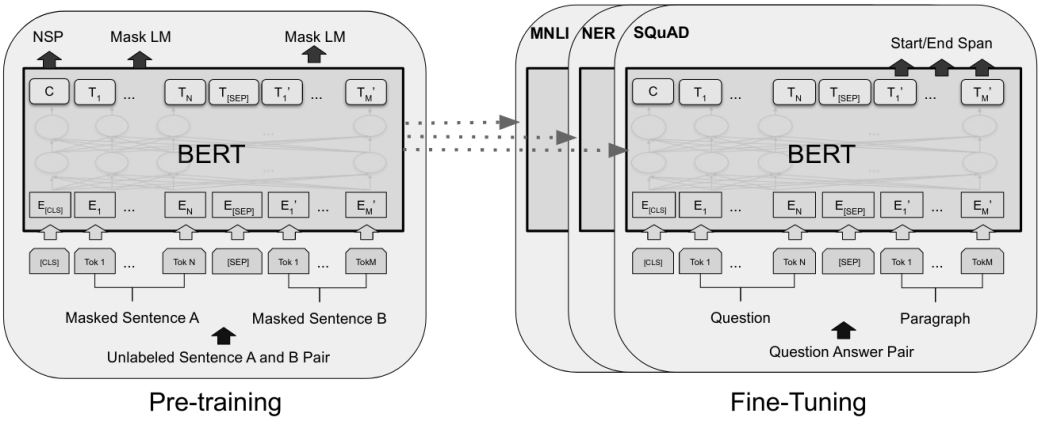

图 1 BERT 模型结构

BERT（Bidirectional Encoder Representations from Transformers）模型是近年来该领域的一项重大研究成果，该模型的出现有效解决了传统训练任务中人工标注任务量大的问题，极大减轻了研究人员的工作量。并为

日后许多研究奠定了坚实的基础。BERT 模型的出现使得我们能够在许多自然语言处理任务中应用一个预先训练的语言模型，缓解了实验人员因数据量或硬件计算能力不足造成的困难。Weijie Liu（2020）结合了知识图谱与 BERT 模型构建了 K-BERT 模型，使其在不同专业领域的非结构化文本的语句分类任务、问答任务与实体命名任务中取得了良好的结果[1]。许雪晨与田侃（2021）提出了 SA-BERT-LSTM 模型并应用于情绪分类，将该模型的输出结果作为股指预测模型的解释变量，结果表明该方法在个股收益的预测精度上达到了 66.41% 的平均准确率[5]。

2 基于先验知识的股价预测模型

2.2 数据选取

本文选用了上证指数 2000 年初至 2022 年初 22 年间的股票历史数据，包括开盘价、收盘价、最高价、最低价、成交数进行了学习，以及所能搜集到的部分同期市场新闻，并以此预测股价的走势。

新闻数据结构如下表所示

表 1 新闻文本数据

NewsDate	NewsTitle	ClassifyName
27/02/2018	浙江义乌集贸市场 2017 年总成交额增长 8.9%	行业新闻
27/02/20186	从合作到反目：科菲特前后实控人“交手”辉丰股份子公司“内讧”难了	金融市场
27/02/2018	欧股小幅收涨关注欧洲经济数据与鲍威尔讲话	外汇新闻

在预测方法方面，我们将前 6 日的历史数据即经过自然语言模型处理的新闻作为输入信号，以此预测股价在第 7 日的表现。注意到，在选取数据时，应当注意数据之间的关系，注意其是否满足独立不相关等条件。同时，为了满足预测的精准性，新闻数据的来源应当全面、客观，否则可能导致模型学习过程中的偏差。

2.3 模型结构

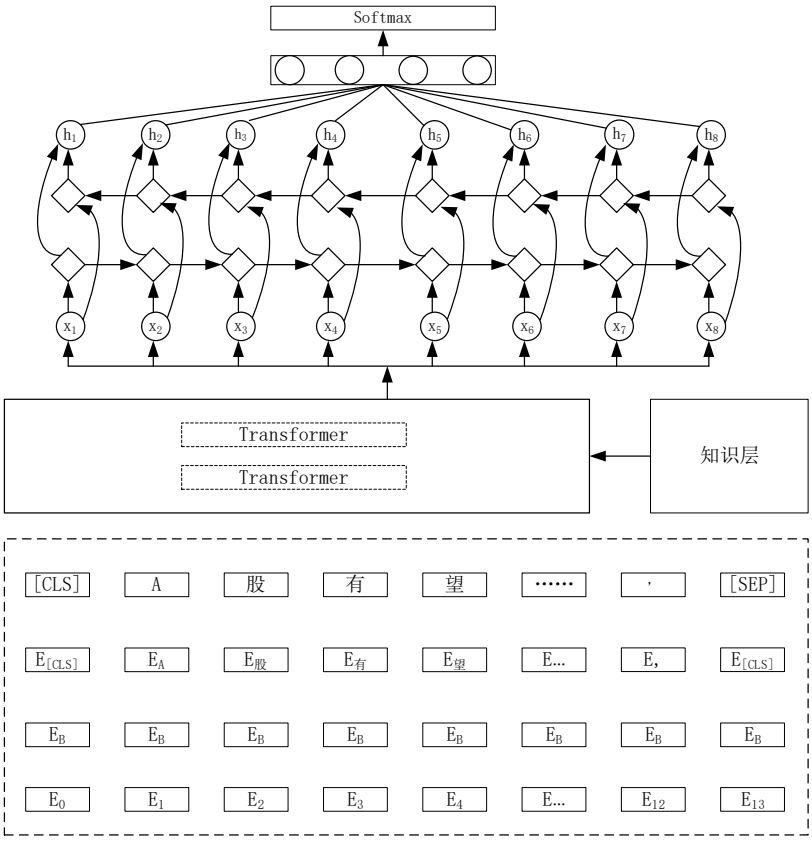

图 3 K-BERT-BiLSTM 模型结构

模型分为情感分析和预测两部分。首先财经新闻标题经过分词后输入 Encoder 编码模块得到对应的词序列转化后的索引，且将文本的最大长度限定为 126 个字，对长度超过 126 的文本进行截断，少于 136 的文本用 0 填充，同时在输入文本的开头和结尾分别添加[CLS]和[SEP]标识符。在得到上述文本的词向量后，将其输入 K-BERT 模型该模型的知识层会提取各语句的主语，并保存其相关关系，当下次预见该主语时，模型的学习会考虑这些已经被学习到的关系。同时，进行 Masked LM 和 NSP 两个预训练任务，从海量文本数据中学习字符级、词语级、语句级和语句间的特征。传统的 BERT 模型一般在大规模语料库上进行训练，然后在下游具体任务上由使用者进行微调。这种方法的不利之处在于处理非结构化文本时，可能会降低模型的效果。因此，Weijie Liu (2020) 提出了 KBERT 模型，在模型 fine-tune 阶段，通过知识图谱引入了外部知识，显著性地提高了 BERT 模型处理非结构性文本的能力。对于本文来说，BERT 在金融文本上预训练可按如下方法进行。

我们将 BERT 的输入中字符[CLS]对应的输出 C 乘以权重 W，作为 Bi-LSTM 网络的输入，计算公式为

$$a_i = g(W_a C + b)$$

然后，模型在隐层中带入输入变量，Bi-LSTM 在两个不同方向的隐层上进行计算，最终将两个方向的结果拼接输出，即$h_i = \overrightarrow{h_i} + \overleftarrow{h_i}$。$\overrightarrow{h}$表示前向传播隐层向量,$\overleftarrow{h}$为后向传播隐层向量。tanh 作为隐藏层的激活函数，计算过程为

$$h_i^d = \sigma(W_h^d a_i + U h_{i-1}^d + b_h^d$$

其中，W_h^d代表第 d 个索引对应的a_i的权重矩阵，U 是隐层在 i-1 时刻的输出h_i^d对应的权重矩阵，d 代表隐藏层

的两个不同方向，b_h^d对应于第 d 个索引向量的偏移量。 最后，将最后一层的所有向量h_i^d拼接起来作为整个句子的特征向量表示

为了对财经新闻文本的情感进行分类，将 Bi-LSTM 输出的特征向量输入 softmax 函数来预测情感分类结果。 分类预测结果如下式所示。 模型会对于每条金融文本输出一个向量，表示该文本是正负面的概率。

$$p(y|H, W_c, b_c) = softmax(W_c H + b_c)$$

此外，本文采用交叉熵损失函数作为目标函数，以最小化目标函数的损失值。

$$Loss = y_i \log(p_i) + (1 - y_i)\log(1 - p_i)$$

时效性在股票市场预测中非常重要，LSTM 模型在处理时间序列数据时具有良好的性能。 因此，使用 LSTM 模型来预测下一个交易日股指价格的涨跌。LSTM 网络有两层。 第一个隐层为时序 LSTM，共有 64 个单元。 输入为 t 交易日市场数据与 t 交易日情绪分析特征的拼接向量； 第二个隐藏层是无时间序列 LSTM。 ，共有 32 个单元。 LSTM 学习到的特征向量先经过全连接层，然后 sigmoid 层输出下一个交易日股价上涨或下跌的对应概率。 sigmoid 函数是一个单调递增的平滑函数，可以将全连接层的输出映射在 0 到 1 之间。

$$h_i = ReLU(W_h[T_t : S_t])$$

$$y_i = sigmoid(W f_c(h_i) + b)$$

模型训练的损失函数为：

$$L(\vec{p}, \vec{d}; \vec{\theta}) = -\sum_{i=1}^{T} [c_{i+1} ln\hat{y}_{i+1} + (1 - c_{i+1}) \ln(1 - \hat{y}_{i+1})] + \lambda \|\theta\|_F^2$$

3 实验设计

本研究基于 Keras 深度学习库，搭建了基于 K-BERT 的神经网络模型对股价进行了预测，预测结果如下图所示

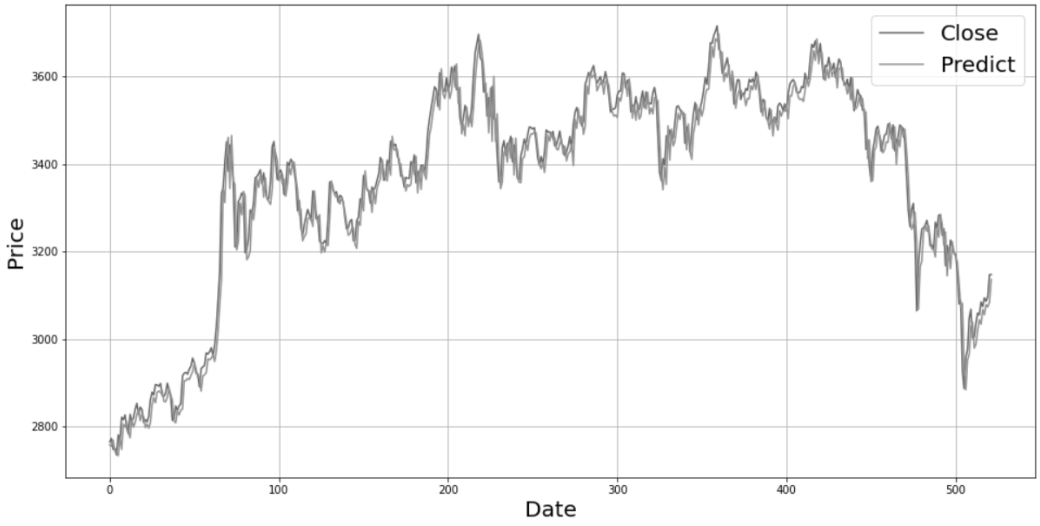

图 3 双向 LSTM 神经网络预测结果

4 模型实现

本研究基于 Keras 深度学习库，搭建了基于 K-BERT 的神经网络模型对股价进行了预测，使用 Python 的 sklearn 与 keras 机器学习库实现情感分析与预测模型。以及 matplotlib 使预测结果可视化。

首先从数据库中提取上证指数自 2000 年以来的所有历史数据，并对其进行划分，以 2021 年 1 月之前的所有数据作为神经网络模型的训练集（training set），自 2021 年之后至 2021 年 12 月的所有历史数据作为验证集（valid set），以验证我们的训练结果的有效性。

其次，对数据进行归一化（Normalisation）处理，这里的目的是为了在预处理数据时使所有数据都落在相同的区间，从而避免一些极端值（奇异样本数据）对预测结果的影响。在最后，再进行一次反归一化，使数据恢复原本的取值。

下面构建 LSTMs 模型以及训练集与验证集。我们首先以序贯模型[1]（Sequential）的结构组织两层 LSTMs 神经网络和一层 Softmax 全连接层。下一步，选择训练模式，我们使用平均平方误差函数和决定系数对模型的预测结果进行评估，其中，平均平方误差越小，模型的预测结果与真实结果的差距越小。决定系数越接近 1，说明模型拟合优度越大。公式如下：

$$R^2 = \frac{\sum_{n-1}^{N}(y_p - y_{avg})^2}{\sum_{n-1}^{N}(y_n - y_{avg})^2}$$

$$MAE = \frac{1}{N}\sum_{n-1}^{N}|(y_p - y_n)|^2$$

然后对验证集数据以及训练数据进行反归一化（inverse-normalization），并使用我们组织好的模型进行训练。

最后，使用 matplotlib 对模型的预测结果与提取数据的收盘价进行绘图，横轴表示了时间，最小间隔为一天，纵轴表示收盘价格，单位为人民币，预测结果如下图所示，预测结果R^2达 0.92，MAE 为 27 符合我们的预期。

5 结论

本研究基于 K-BERT 模型，改进了其预测股价的模块，利用 Keras 深度学习工具箱搭建了股价预测模型，对上证指数走势进行了预测，并获得了合意的结果。从拟合结果来看，尽管存在着时间上的偏差，但我们的模型较为近似地预计了上证指数的价格趋势。此外，在这个模型中也存在着一些问题：首先就数据集的大小而言，从 2000 年开始至今也不过仅有 5000 条数据，这样的数据量由于数据过少，导致该模型泛化能力不足，若希望预测其他股指的走势，则必须使用其他股指的历史数据重新训练。

但从模型本身的优点来看，LSTMs 模型本身非常适用于处理一些带有时序信息的数据，在信息高度交互的今天，许多消息可以通过互联网新闻，媒体以及社交网站快速传播，从而影响在投资者将会做出的决策。而这些信息对于预测股价来说是不可或缺的。而引入了先验知识的 BERT 模型则可以很有效地处理一些非结构化的文本数据，分析其情感内容并以此预测投资者对该信息的反应。

本文的结论反映了深度学习模型在现代金融经济学研究中的重要作用，深度学习模型以其强大的数据

[1] 序贯结构是最简单的线性结构顺序。

处理能力将会对金融研究，尤其是资产定价领域的理论起到重要的推进作用。因为在有效市场理论的支持者看来，任意时刻的金融资产的信息都体现在了价格上，当我们具备一个足够强大的计算机和足够多的数据时，是否意味着我们能在任意时刻对任何一个资产实时的做出定价，并以此推进金融市场向有效的方向发展？这在将来的金融学研究中将会是一个值得被探讨的问题。

参考文献：

[1] LIU W, ZHOU P, ZHAO Z, et al. K-BERT: Enabling Language Representation with Knowledge Graph[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(3): 2901–2908. DOI:10.1609/aaai.v34i03.5681.

[2] 姜富伟, 孟令超, 唐国豪. 媒体文本情绪与股票回报预测[J]. 经济学(季刊), 2021(4): 1323–1344.

[3] HOCHREITER S, SCHMIDHUBER J. Long Short-Term Memory[J]. Neural Computation, 1997, 9(8): 1735–1780. DOI:10.1162/neco.1997.9.8.1735.

[4] LIU L, ZHANG T. Economic Policy Uncertainty and Stock Market Volatility[J]. Finance Research Letters, 2015, 15: 99–105. DOI:10.1016/j.frl.2015.08.009.

[5] 许雪晨,田侃.一种基于金融文本情感分析的股票指数预测新方法[J].数量经济技术经济研究,2021,38(12):124-145.DOI:10.13653/j.cnki.jqte.2021.12.009.

[6] 徐浩然,许波,徐可文.机器学习在股票预测中的应用综述[J].计算机工程与应用,2020,56(12):19-24.

[7] 周学伟,付巾书,宋加山.不同的政策不确定性对股市波动影响相同吗？[J].金融发展研究,2020(05):78-85.DOI:10.19647/j.cnki.37-1462/f.2020.05.013.

[8] 杨春鹏.基于展望理论的证券市场反应过度和反应不足研究[J].运筹与管理,2007(06):118-122.

[9] PANG B, LEE L, VAITHYANATHAN S. Thumbs Up?[C]//Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - EMNLP '02. . DOI:10.3115/1118693.1118704.

[10] 黄虹,卢佳豪,黄静.经济政策不确定性对企业投资的影响——基于投资者情绪的中介效应[J].中国软科学,2021(04):120-128.

[11] 王琳,刘宏雅,沈沛龙.央行政策沟通的金融资产价格稳定效应——以中国股票市场为例[J].金融论坛,2021,26(08):8-17.DOI:10.16529/j.cnki.11-4613/f.2021.08.003.

[12] MALKIEL B G, FAMA E F. EFFICIENT CAPITAL MARKETS: A REVIEW OF THEORY AND EMPIRICAL WORK*[J]. The Journal of Finance, 1970, 25(2): 383–417. DOI:10.1111/j.1540-6261.1970.tb00518.x.

[13] Jay Ritter. Behavioral finance[J]. Pacific-Basin Finance Journal, 11(4):429–437, 2003.

[14] LIU L, ZHANG T. Economic Policy Uncertainty and Stock Market Volatility[J]. Finance Research Letters, 2015, 15: 99–105. DOI:10.1016/j.frl.2015.08.009.