基于多维异构视角的多分支 LSTM 模型对股票市场系统性风险的预测研究——以中国 A 股为例

兰文婕

摘要

中国把防范化解系统性金融风险放在重要位置，因此如何有效地预测系统性金融风险成为了中国式现代化背景下的一项重要议题。为此，本文提出了一种由多模态数据驱动的多分支 LSTM 模型进行中国 A 股市场系统性风险预测。

本文通过多维异构的数据驱动，进行模型的指标选取：传染效应、文本和基础指标。对于传染效应指标的构建，本文通过收集全球不同国家（地区）的市场指数，提取市场间风险结构关联图中的特征和基于 CoES 构建的相关性向量。收集 A 股市场相关的财经新闻文本，提取文本信息作为文本指标。通过宏观经济、股票市场、外汇三个层面的考虑收集基础指标，也包括了系统性风险自身的历史信息。三个异构的不同模态数据提取出的指标共同构成本文模型学习的指标特征。

接下来，本文创新性地构建了多分支 LSTM 模型用于预测 A 股市场的系统性风险。对于不同模态数据得到的指标特征，本文采用了三个各自独立的 LSTM 模型分支进行训练，保证信息的有效提取。同时，考虑到指标特征整体的信息及内在联系，本文单独采用了一个卷积层及 LSTM 模型构成的分支，用于训练模型学习整体知识。

研究结果表明，本文构建的传染性网络特征有效提高了模型的有效性，在监控预测风险时可以着重关注。同时本文构建的多分支 LSTM 能够有效学习到知识，有助于辅助监测预警股票市场的系统性风险。

关键词：系统性风险；多维异构数据；传染效应；文本信息抽取；多分支 LSTM 模型

一、绪论

（一）选题背景与意义

如何有效地监测系统性金融风险，并基于监测结果针对性地防范化解风险，是中国式现代化背景下的一项重要议题。金融市场系统性风险的成因包括多个方面，以往的研究一般通过股市收益率、债市利率等基础性宏观经济指标构建。为扩展现有数据，拓展多维异构数据，本文引入市场之间系统性风险传染性网络特征和文本信息特征。其中，随着全球化进程的加快，各个国家地区之间的股市联动性进一步扩大，单个市场的风险极易通过风险传染网络传导到其它地区，A 股市场极易受到其它地区股市的影响，因此通过市场之间系统性风险传染性网络提取的特征也能够扩大风险的预测源；同时，大数据时代下，文本信息也成为重要的信息源之一。

本文以中国式现代化下的该议题为背景，基于股票市场中的系统性金融风险隐患，建立相应的深度学习模型，对风险进行预测，为我国现代化建设过程中防范化解重大风险提供科学参考，进而增强我国防范化解重大风险的能力，推进国家治理体系治理能力现代化。

（二）文献综述

学界与业界十分关注如何监测风险、预测风险。针对于系统性金融风险的预测，学术界也给出了较多的方法。传统的预测模型基于金融市场数据的时序特征，多用传统回归分析与时间序列模型进行预测（Valaskova 等, 2018）。除了基本的时序模型之外，更为广泛运用的是自回归模型（Taylor 和 Yu, 2016）。同时随着金融风险管理学科的发展，VaR 等模型也应用于风险预测与管理。Gerlach 等（2011）的研究则根据贝叶斯推论，结合传统 VaR 模型提出了一个精度更高的风险预测模型。近年来，数据科学领域的机器学习、深度学习模型也逐渐应用于金融风险预测领域，如分类算法的优化与运用(Peng 等, 2019)、多任务监督模型的应用(Sawhney 等, 2020)以及词向量和文本回归方法的应用(Yeh 等, 2020)。

二、指标构建与数据预处理

（一）多维异构数据来源

本文在 CSMAR、CEI 等数据平台上收集了 A 股市场与系统性风险有关的多维异构的数据，即类型结构和特征不一致的数据，将时间范围限制在 2018 年至 2022 年，同时将非交易日的数据剔除，总共得到 1215 个交易日的时间序列数据。

数据集包含三个方面的异构数据：用于计算传染效应特征的多国市场的指数数据、用于提取文本特征的交易日 A 股市场财经新闻文本数据与股票市场系统性风险相关的宏观经济、股票市场和外汇方面的基础指标数据。

（二）基于 CoES 的传染效应指标构建

为了准确衡量各个市场之间的传染性关系，本文提取 2017 年至 2022 年，包括中国、美国、日本、德国等共计 12 个国家（地区）的具有代表性的指数，对应指数如表所示：

表 1 市场指数表

国家（地区）	指数	国家（地区）	指数
中国	富时中国 A50	英国	富时 100
中国香港	恒生指数	法国	CAC40
中国台湾	台湾加权指数	韩国	KOSPI
美国	标准普尔 500	日本	日经 225
印度	孟买 30	欧洲	斯托克 50
俄罗斯	俄罗斯交易系统市值加权指数	德国	DAX30

本文基于上述指数构建出每个市场的对数收益率$R_{(i,t)}$，$i \in [1,N]$，$t \in [2,T]$。基于计算出的对数收益率，本文进一步根据 Tobias 和 Brunnermeier（2016）提出的条件期望损失（CoES）衡量单个市场在特定条件下的期望损失。在市场 j 处在α水平下的困境时，市场 i 的条件期望损失定义为：

$$CoES_{ij,t}(\alpha) = E\left[R_{i,t} \mid R_{j,t} < VaR_{j,t}(\alpha)\right] \qquad \text{公式（1）}$$

其中，$R_{i,t}$是市场i在 t 时刻的收益率，$VaR_{j,t}$是市场 j 在 t 时刻的风险价值。为衡量不同市场之间的关联，本文参考 Chen（2019）的研究，构建 12 个市场之间的风险结构关联网络。在每个时间节点 t，每个市场都可得到一个维度为 12 维的风险结构向量$X_{i,t} = \left\{CoES_{ij,t}\right\}_{j=1,\cdots,N}$。再通过计算各个市场的余弦相似度来衡量极端事件下不同市场期望收益变化的相似程度：

$$\rho_{ij,t} = \frac{X_{i,t}^{T} X_{j,t}}{\|X_{i,t}\|\|X_{j,t}\|}, j \neq i, i = 1,\cdots,N, t = 1,\cdots,T. \qquad \text{公式（2）}$$

本文得到 1215×12×12 维度的数据，其中 1215 为 2018 年至 2022 年的交易

日时间长度，本文为每个交易日构建出 12 个市场之间的风险结构关联网络，为了提取风险结构网络中的特征，分两种方式提取特征：

方法一：提取 A 股市场与其他市场的基于 CoES 构建的相关性向量，维度为 11 维。其余 11 个市场的风险波动通过相关性直接传导至 A 股市场，基于此获取直接传染性特征。

方法二：利用自编码算法（AutoEncoder）从图中提取特征，算法包括两个核心部分：编码器与解码器。首先本文将高维输入编码为低维度的隐变量，解码器部分再将重构为给定维度，在此处本文将实验 1、4、8、16 维，选取效果最佳的维度，原理如下公式所示：

$$s = g_{\theta_1}(c_t) = \sigma(W_1' hi_t + b_1) \qquad \text{公式（3）}$$

$$\widehat{c_t} = g_{\theta_2}(s) = \sigma(W_2 s + b_2) \qquad \text{公式（4）}$$

其中，c_t是未经过处理的相关性矩阵特征，$\widehat{c_t}$是经过自编码器编码后的相关性矩阵特征。通过这一方式，除方法一构建直接传染性特征外，本文可以进一步衡量网络中复杂的传染效应。

（三）基于 FinBERT 的新闻文本指标构建

对于文本指标特征的构建，本文获取了 2018 年至 2022 年中国 A 股市场所有交易日的财经新闻文本数据，采用财经新闻的新闻标题代表整篇新闻的内容，数据结构示意如下表：

表 2 新闻文本数据结构表

公布日期	新闻标题
2018-01-01	中国世界工厂时代已过去 外媒：正成为世界消费者
…	…
2018-12-31	林达控股(01041.HK)延长建议发行非上市认股证券达成日期

对于新闻标题的文本向量化处理，本文考虑到在金融领域的大背景下，通用领域表现良好的 BERT 等模型对金融专业词汇的学习不够有效，表现会相应降低，因此使用 Liu 等(2021)提出的 FinBERT 模型对新闻数据进行向量化处理。

对每个交易日 t 的 n 个新闻文本new_t，通过 FinBERT 模型，得到了对应 n 篇新闻的 768 维向量。由于矩阵规模较为庞大，本文对每个交易日的所有特征向量求均值作为当日的文本特征h_t：

$$h_t = \sum_{i=1}^{n} FinBERT(new_t) \qquad \text{公式（5）}$$

接下来，考虑到 768 维向量长度过长可能导致后续 LSTM 模型学习效果，对文本特征进行降维处理，本文采用与传染效应指标构建类似的自编码器进行降维，最终将 768 维的文本特征转换为 11 维的文本特征。

（四）基础指标构建与系统性风险的度量

1. 基础指标的选取与构建

除传染效应和新闻文本指标外，本文也将系统性风险的历史数据也当作可以学习的特征放入模型当中，将其纳入基础指标的范畴。

本文从宏观经济、股票市场、外汇三个方面进行基础指标的选取。其中，对月度数据和季度数据，本文将其填充赋值转化为日度数据。指标如下表所示：

表 3 系统性风险基础预测指标

分类	频度	指标名称
宏观经济方面	月度	固定资产投资
	季度	GDP 增长指数
	日度	三年期国债收益率
	季度	工业增加值增长指数
股票市场方面	日度	沪深 300 指数回报率
	日度	深证综合 A 股指数回报率
	日度	中证 100 指数回报率
	日度	上证综合 A 股指数回报率
	日度	综合日市场交易总金额
	日度	综合日市场交易总股数
外汇方面	日度	港币对人民币元汇率
	日度	人民币元对美元汇率

2. 系统性风险的度量

为了衡量中国 A 股市场系统性风险，本文根据富时 50 指数计算对数收益率，再据此计算期望损失（ES）。窗口期选取 2017 年全年交易日数量，即 242 天：

$$ES_{i,t}(\alpha) = E\left[R_{i,t} < VaR_{i,t}(\alpha)\right]$$ 公式（6）

三、描述性统计分析

首先对传染效应指标数据进行分析。本文衡量中国 A 股市场与另外 11 个市场之间的传染性关系，建立了 12 个市场间每个交易日共 1215 个风险结构关联网络。关联网络为 12×12 大小的矩阵：

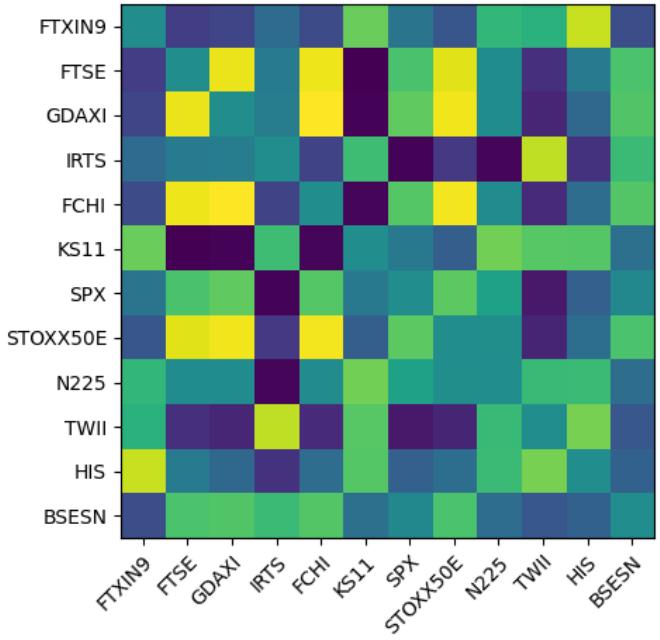


图 1 风险结构关联网络矩阵热力图

上图即是 2022 年 12 月 30 日的风险结构关联网络矩阵，其中当日 HIS、KS11 指标相对来说相关性较高，而 FTSE、GDAXI 等指标则相对较低。

之后通过方法一根据各市场的指数计算 CoES，构建了 1215 个交易日的 11 维时间序列数据作为特征，画出折线图如下：

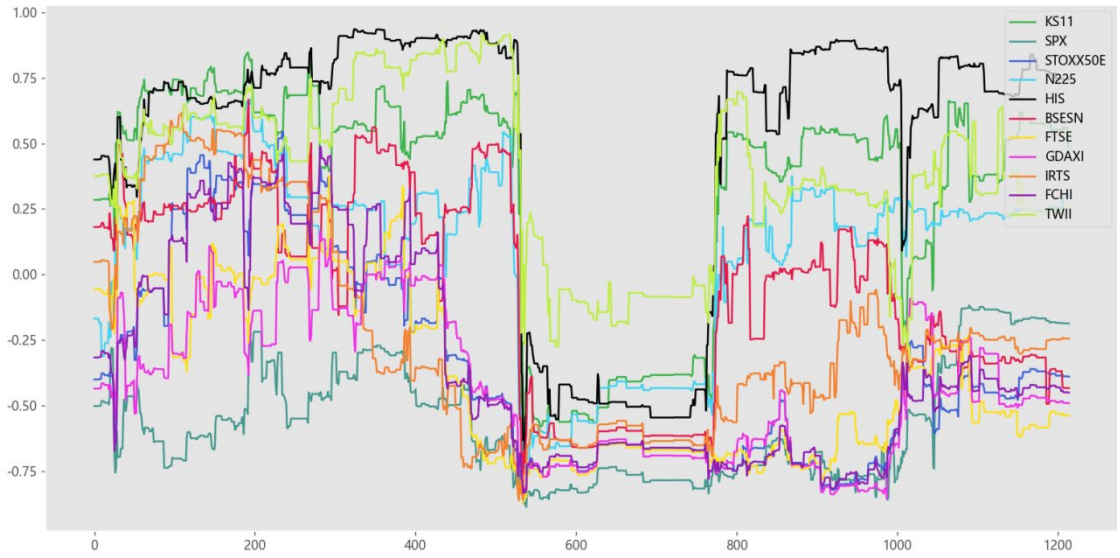
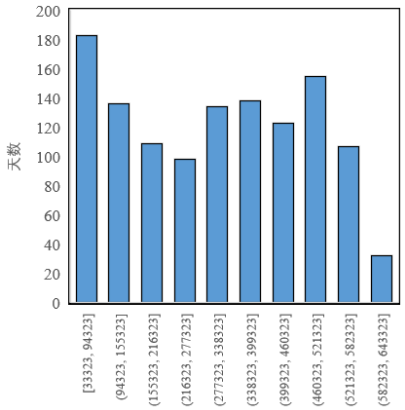


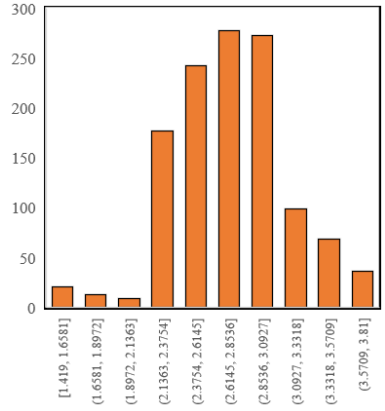
图 2 传染效应指标折线图

可以发现，在第 500 个交易日至第 800 个交易日期间，几乎所有指标都迅速下降并维持在了较低水平，然后重新回升。从图中可以得出，HIS、TWII、KS11 指标与 A 股市场的相似度最高，而 SPX 与 A 股市场的相似度最低。

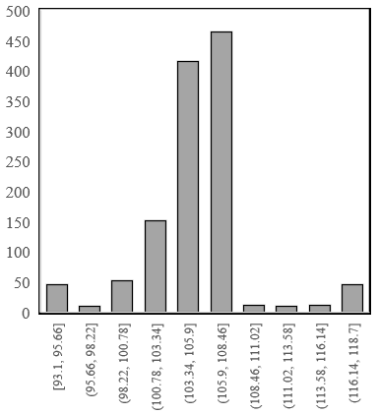
然后对基础指标数据分析，本文将月度的与季度的宏观经济指标扩展为了日度的指标，将当月的每一天都按照月度数据进行填补，画出直方图统计如下：



固定资产投资



三年期国债收益率



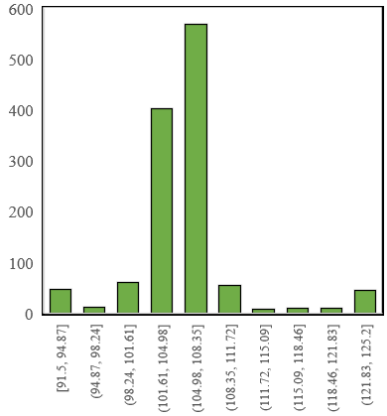


图 3 宏观经济指标统计直方图

四、多分支 LSTM 模型的建立

（一）模型结构设计

本文模型使用 pytorch 进行神经网络的架构。在一般 LSTM 模型的基础上进行了创新性的改进。模型输入特征分为传染效应特征c_t、文本特征h_t和基础特征f_t。本文首先对输入的数据进行归一化处理以保证基本度量单位的统一。本文使用最大最小标准化方法来进行归一化：

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \qquad \text{公式（7）}$$

接下来将数据划分为训练集与测试集，本文设置比例为 9:1，即训练集有 1094 天，测试集有 121 天。

输入数据实际上是三种异构数据得到的不同模态的指标信息。因此，本文考虑使用三个不同的 LSTM 层来分别对这三种数据进行训练，模型时间窗口设置为五天。模型的前向传播完整的输入流程如下，首先分别使用三个不同的 LSTM 模型学习传染效应特征、文本特征和基础特征，得到独立的三个输出：

$$out_t^x = LSTM(x_{t-1}, \dots, x_{t-5}) \ , x = c, h, f \qquad \text{公式（8）}$$

同时，由于三种模态的数据之间可能存在隐藏的关联信息，本文需要设计一个模块用于捕捉整体的数据信息。考虑到 CNN 用于特征提取的良好表现，因此将三类数据拼接合并后通过一层卷积层充分提取信息，再输入 LSTM 层继续捕捉信息：

$$out_t^{concat} = LSTM\big(Conv(x_{t-1}, \dots, x_{t-5})\big)$$
$$x_t = concat(c_t, h_t, f_t) \qquad \text{公式（9）}$$

最终将四个 LSTM 层的模型输出进行拼接，通过全连接层来进行传染性风险的预测。全连接层输出张量的参数设置为 1 以进行最终的数值预测：

$$pred = Softmax\big(concat(out_t^c, out_t^h, out_t^f, out_t^{concat})\big) \qquad \text{公式（10）}$$

多分支 LSTM 模型的结构如下图所示：

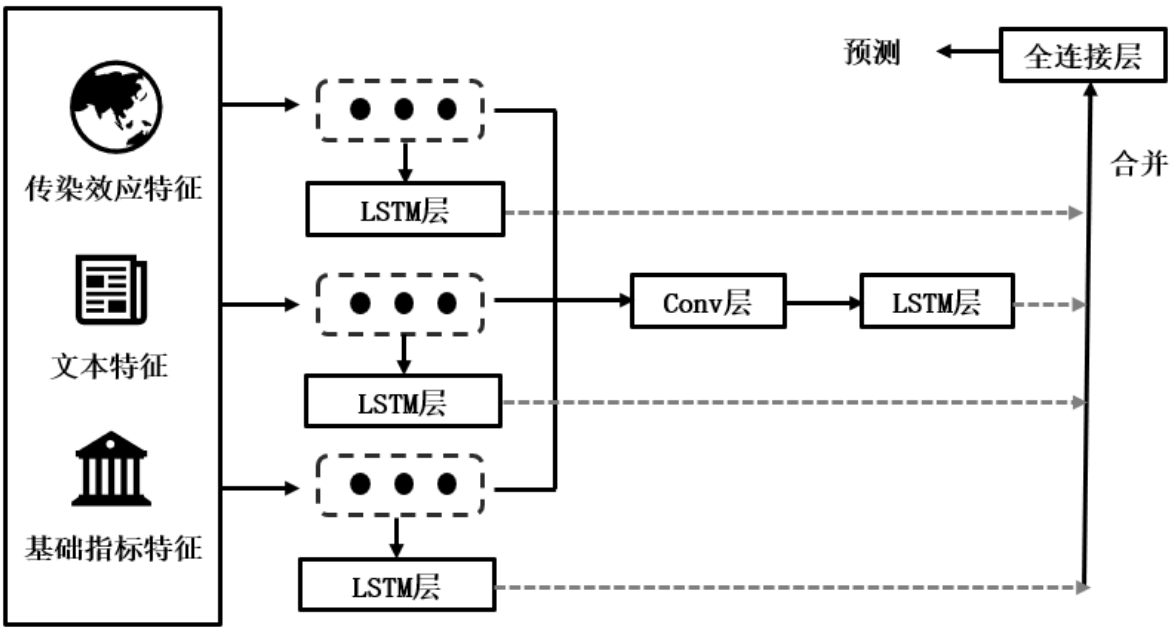


图 9 多分支 LSTM 模型整体框架

其中对于单个 LSTM 模型，本文经过调参尝试，根据模型效果表现，最终设置模型的参数如下：

表 4 模型参数表

参数	值
时间步长	5
隐藏层数	6
隐藏层神经元数	12
批大小	64
学习率	0.001
迭代次数	5000

本文使用 MSELoss 函数作为模型学习的损失函数。MSELoss 函数是通过计算均方误差（MSE）作为损失进行模型训练的。均方误差是反应预测值与被真实值之间差异程度的一种度量方法。设预测值为$\hat{y}$，真实值为y，则 MSELoss 计算损失的公式为：

$$loss(\hat{y}, y) = \frac{1}{n}\sum(\hat{y}_i - y_i)^2 \qquad \text{公式（11）}$$

五、实验结果

（一）系统性风险预测结果

1. 模型评价指标

为评估多分支 LSTM 模型的预测效果，本文使用平均绝对误差（MAE）和均方根误差 （RMSE）作为指标来进行模型评价。MAE 和 RMSE 是关于连续的数值型变量的两个最普遍的度量标准，原理如下：

$$MAE = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n} \qquad \text{公式（12）}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}} \qquad \text{公式（13）}$$

其中，$\hat{y}_i$为预测值，在实验中代表预测出的系统性风险，y_i为实际值。MAE 和 RMSE 值越小，则代表模型的预测效果越好。

2. 预测结果

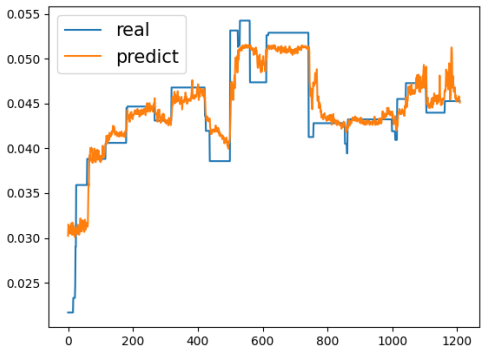
本文对多分支 LSTM 模型进行了多次实验，得到预测结果。对每次实验的测试集的预测结果与真实值计算 MAE 和 RMSE，求平均得到结果如下：

表 5 预测结果表

	MAE	RMSE
训练集	0.0016953943	0.002449348
测试集	0.0014233027	0.0017694924

可以发现，多分支 LSTM 模型的拟合效果在训练集上和测试集上表现良好，MAE 和 RMSE 都较低。其中测试集的拟合效果要略优于训练集。

画出一次实验中的训练集（左）与测试集（右）的拟合曲线图如下：



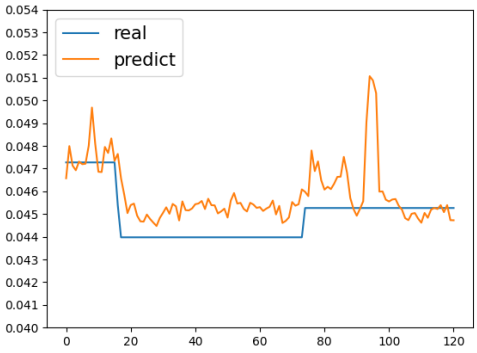


图 10 模型拟合效果图

可以看出，训练集的拟合情况良好。而测试集拟合的预测值和真实值的趋势大致相同，但有一定的波动和异常值出现，总体来说模型的预测表现是有效的，模型的泛化性能较高且鲁棒性较强。

（二）基线对比与消融实验结果

1. 基线模型对比结果

为了证明多分支 LSTM 模型对系统性风险预测的有效性，本文将其与三个基线模型进行比较。模型分别为 GRU 模型、一般 LSTM 模型和 CNN-LSTM 模型。模型对测试集的预测结果表现对比如下：

表 6 与基线模型的对比结果表

模型	MAE	RMSE
LSTM 模型	0.0017645524	0.0018937876
GRU 模型	0.0024787416	0.0026822179
CNN+LSTM 模型	0.0023522072	0.0025513992
多分支 LSTM 模型	**0.0012573128**	**0.00161096435**

从表中本文可以观察到：与其他基线模型相比，多分支 LSTM 模型在训练集和预测集上都取得了最好的结果，这说明了多分支 LSTM 模型对股票市场系统性风险预测的有效性。此外，优于 CNN-LSTM 模型表明了本文的设计可以很好地捕捉到三种模态数据的信息及内在联系。同时，优于 GRU 模型和 LSTM 模型则说明了改进的模型结构的合理性。

2. 消融实验结果

为验证本文选取指标的有效性和模型设计的合理性，本文进行了多次不同的消融实验。

首先是指标选取的有效性，本文指标划分为三部分：传染效应指标、文本指标和基础指标。依次对三部分指标两两组合，构建新数据集进行实验。此时的多分支 LSTM 模型由于只学习了两个模态的数据，因此只有两个分支。得到结果如表 7：

表 7 不同指标组合的测试集预测结果表

指标组合	MAE	RMSE
传染效应指标+文本指标	0.001851739	0.0021131611
传染效应指标+基础指标	0.0014377916	0.0016692132
文本指标+基础指标	0.0014557169	0.0017422976
所有指标	0.0012573128	0.00161096435

画出条形图进行进一步可视化对比：

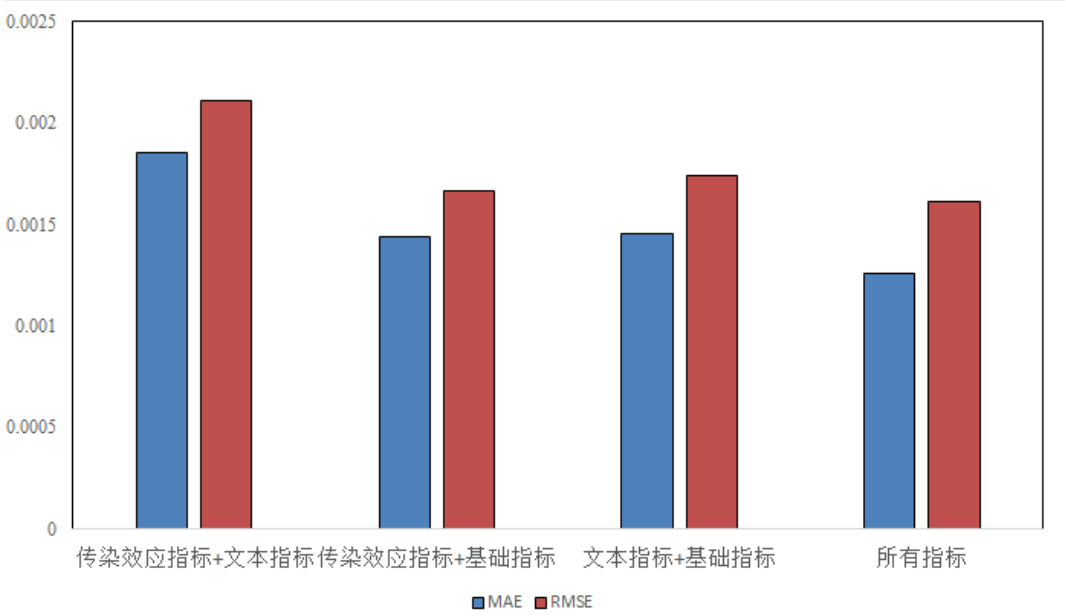


图 11 不同指标组合 MAE 与 RMSE 条形图

可以发现采用所有指标数据进行学习的多分支 LSTM 模型相对于其他的组合来说取得了最好的效果，因此可以证明本文三部分指标选取的有效性。同时可以发现只有文本和传染效应指标时训练出的多分支 LSTM 模型的 MAE 与 RMSE 最大，因此可以推断出基础指标对于预测市场系统性风险是较为有效的。除此之外，可以发现文本指标和基础指标的指标组合仅是略逊于所有指标，所以文本指标中所能学习到的知识是较少的。

六、结论与建议

为了提高金融市场系统风险的预测准确性，通过预测指标的警示作用增强金融市场的抗风险力，本文综合以往的研究成果构建了三类风险源：基于 CoES 构建的传染性网络特征、基于 FinBert 构建的新闻文本特征、以及大盘指数等一系

列基础指标特征的信息，创新性地构建了多分支 LSTM 算法，有效地提取了三类风险源的信息，模型具有较强的预测性。

从特征源效果看，本文构建的传染性网络特征有效提高了模型的有效性，这一结论也可以从经济学原理得到验证，以往研究大量表明中国 A 股市场与其它股票市场之间有着强烈的联动性。通过构建传染性矩阵，本文将 A 股市场置于全球股市关系网络下，衡量它受到外部冲击后自身极端风险的变化水平。新闻文本信息效果不佳的原因可能有如下几点：第一，新闻文本样本质量较差，金融市场相关信息量较少，导致训练效果较差；第二，自编码未能有效提取的文本信息。基础性指标的预测性效果最好，大量过往研究已证实这类指标为市场风险源。

从模型效果看，本文构建的多分支 LSTM 模型有效地提取了多维异构数据得到的特征中的有效信息，不仅考虑了每种特征由于异构数据源而保留的独立信息，通过添加卷积层进一步结合了三类特征之间的相关关系，从而提高预测的准确性。

参考文献

[1] Adrian, T., & Brunnermeier, M. K. (2016). CoVaR. THE AMERICAN ECONOMIC REVIEW.

[2] Bai, S., Kolter, J. Z., & Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling.

[3] Chen, C. Y. H., Härdle, W. K., & Okhrin, Y. (2019). Tail event driven networks of SIFIs. Journal of Econometrics, 208(1), 282-298.

[4] Gerlach, R., Chen, C. W., Lin, E. M., & Lee, W. (2011). Bayesian Forecasting for Financial Risk Management, Pre and Post the Global Financial Crisis. OME WORKING PAPER SERIES.

[5] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

[6] Grilli, R., Tedeschi, G., & Gallegati, M. (2014). Markets connectivity and financial contagion. Journal of Economic Interaction and Coordination, 10(2), 287-304.

[7] Liu, Z., Huang, D., Huang, K., Li, Z., & Zhao, J. (2021). Finbert: A pre-trained financial language representation model for financial text mining. In Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence (pp. 4513-4519).

[8] Peng, Y., Wang, G., Kou, G., & Shi, Y. (2011). An empirical study of classification algorithm evaluation for financial risk prediction. Applied Soft Computing, 11(2), 2906-2915.

[9] Sawhney, R., Mathur, P., Mangal, A., Khanna, P., Shah, R. R., & Zimmermann, R. (2020). Multimodal Multi-Task Financial Risk Forecasting. Paper presented at the Proceedings of the 28th ACM International Conference on Multimedia.

[10] Taylor, J. W., & Yu, K. (2016). Using Autoregressive Logit Models to Forecast the Exceedance Probability for Financial Risk Management. Journal of the Royal Statistical Society, 179. 、

[11] Valaskova, K., Kliestik, T., Svabova, L., & Adamko, P. (2018). Financial Risk Measurement and Prediction Modelling for Sustainable Development of Business Entities Using Regression Analysis. Sustainability, 10(7).

[12] Yeh, H.-Y., Yeh, Y.-C., & Shen, D.-B. (2020). Word Vector Models Approach to Text Regression of Financial Risk Prediction. Symmetry, 12(1).

附录

市场指数与英文代码对应表：

指数	代码	指数	代码
富时中国 A50	FTXIN9	富时 100	FTSE
恒生指数	HIS	CAC40	FCHI
台湾加权指数	TWII	KOSPI	KS11
标准普尔 500	SPX	日经 225	N225
孟买 30	BSESN	斯托克 50	STOXX50E
俄罗斯交易系统市值加权指数	IRTS	DAX30	GDAXI