基于嵌入和标签平滑，利用新闻文章改进股价走势分类

Luis Villamil
villamil@cs.fsu.edu

Shaeke Salman
salman@cs.fsu.edu

Conrad Horn
horn@cs.fsu.edu

Ryan Bausback
rwb16@my.fsu.edu

Ting L. Liu
tinglliu@bu.edu

Xiuwen Liu
liux@cs.fsu.edu

arXiv:2301.10458v1 [cs.LG] 25 Jan 2023

摘要——**股价走势预测是金融领域一项极具挑战且至关重要的问题。现代行为金融学已充分证实，相关股票的股价往往会在新闻发布后，随着投资者的反应及过度反应而变动。然而，如何通过定量模型捕捉价格变动与新闻文章之间的关系，仍是一个活跃的研究领域；现有模型已取得不同程度的成功。本文提出，通过引入深度学习中的正则化和优化技术，利用新闻文章改进股价走势分类。具体而言，我们采用近期模型中的嵌入和双向循环神经网络，捕捉新闻文章与股票之间的依赖关系。我们进一步引入权重衰减、批量归一化、Dropout 和标签平滑，以提升训练模型的泛化能力。为应对批量归一化导致的验证准确率大幅波动问题，我们提出双阶段训练，以可靠地实现性能提升。在常用数据集上的实验结果表明，模型取得了显著提升：测试集平均准确率达到 80.7%，较现有模型绝对提升超过 10.0%。消融研究表明，批量归一化和标签平滑效果最为显著，平均分别带来 6.0% 和 3.4% 的绝对提升。**

一、引言

现代金融的基础原则之一是有效市场假说（EMH）。该理论认为，在有效市场中，股价会充分反映所有可用的相关信息 [1]。这给任何试图预测股票价格的人带来了难题，因为预测所需的所有信息都已反映在价格中，因此无法可靠地获得任何金融优势[2]。

尽管如此，已有研究在多个情境中表明，基于文本的信息，例如 Twitter [3],新闻文章 [4],甚至 Reddit [5], 都可用于准确预测未来的价格变动。例如，Bollen 等人利用 Twitter [5]预测股票价格的涨跌，取得了 87.6% 的准确率。只需看看 Reddit 上近期的“r/WallStreetBets”事件，就能发现一个例子：来自基于文本的互联网资源的情绪在未来价格变动中发挥了重要作用。

另一个例子是，Li 等人 [3] 描述了一条关于巴拉克·奥巴马受伤的 2013年 虚假推文，该推文导致道琼斯工业平均指数在两分钟内下跌 100 点。因此，显然股票市场确实会对新闻文章等媒体作出反应。

然而，要捕捉这两个系统之间复杂的依赖关系，是一项艰巨的任务。这样的模型不仅需要高效地计算相关的新闻文章特征，还需要能够很好地推广到所有股票和新闻文章。一种表示这些复杂依赖关系的方法是采用自然语言处理 （NLP）技术 [6]。近年来，这类技术取得了革命性进展，在某些情况下，神经网络在相同任务上的表现甚至超过了人类。例如，在斯坦福SQuAD 2.0阅读理解评测中，自然语言处理的F1分数一贯高达93.2% [7]。相比之下，人类仅达到89.4%。

本文进一步基于 Du 和 Tanaka-Ishii 的研究 [6]，展开。他们使用此类自然语言处理技术，将股价和新闻文章转化为嵌入，随后利用门控循环单元（GRU）神经网络，将这些嵌入应用于判断股价将上涨还是下跌的二元分类问题。我们进一步研究批量归一化、Dropout、权重衰减和标签平滑等深度学习技术，如何在相同设置下提升双向 GRU（Bi-GRU）神经网络的性能。首先，我们详细介绍分类问题，以及如何通过 Word2Vec 和 BERT 这两种自然语言处理技术生成嵌入。然后，我们说明数据集、我们提出的全新先进训练方法，以及模型在实验中于该数据集上的表现。我们还简要讨论模型在预测问题上的表现。最后，我们探讨模型可能存在的问题及其影响，并提出若干未来的改进方向。

II. 相关工作

股票价格走势预测已成为自然语言处理研究人员日益关注的研究课题，他们致力于

理解金融市场的波动行为。随着自然语言处理技术的最新发展，自动情感分析的准确度已大幅提升。

已有多项研究提出了多种尝试基于情感预测股价变动的技术。例如，Bollen 等人利用 OpinionFinder[8] 来识别推文中的正面和负面情感。他们还使用 GPOMS 衡量同一批推文所反映的六种不同人类情绪，以预测道琼斯工业平均指数（DJIA） [5]的收盘值。虽然这类特征可用于对文本情感进行分类，但如今也可以从新闻文章及其他基于文本的内容中提取有用信息，并将其表示为连续向量。更正式地说，这些固定长度的向量称为文本嵌入。这些嵌入向量通常包含从上下文中提取的丰富语义信息，因此能够进行数值计算 [9]。

Word2Vec 是一种用于训练词嵌入的算法。这些嵌入通过为词语生成数学表示，帮助定义词语的含义。随后，可利用点积或余弦相似度等相似度函数，通过这种表示捕捉词语之间的关系。Word2Vec 有两种算法：Skip-gram 和连续词袋（CBOW） [10]。在 Skip-gram 模型中，它使用一组句子（语料库）来预测输入词的邻近词（或上下文）。在 CBOW 方法中，它根据上下文词预测目标词。Word2Vec 的引入推动了针对序列开发新型自然语言处理模型，例如 ELMo；它是一种双向 LSTM [11],应用，采用长短期记忆（LSTM）或门控循环单元（GRU）等循环神经网络（RNN）。LSTM 是一种使用输入门和输出门来解决梯度消失问题的 RNN [12]。GRU 是另一种与 LSTM 类似的 RNN，但只有两个门，即重置门和更新门，从而简化并加快训练过程 [13]。

随着Transformer近期的引入，许多曾被认为颇具挑战性的任务如今变得更容易实现。例如，在过去五年中，采用Transformer的自然语言处理方法在斯坦福SQuAD数据集 [7]的阅读理解问题上，在精确率与召回率（F1值）和精确匹配（EM）方面已超越人类。表现优于人类的大多数方法都使用一种名为BERT的Transformer系统。基于Transformer的双向编码器表示（BERT） [14], 是一种利用Transformer模型 [15] 对整段文本进行分类，而非逐词分类的技术。这意味着BERT输出的向量经过编码，能够表示句子的上下文。因此，利用Transformer可以获得融入上下文的句意表示，使模型更容易学习。

Du 和 Tanaka-Ishii [6] 基于 Hu 等人 [16], 的研究开展工作，是最早将注意力机制应用于新闻驱动的股价变动分类的研究之一。Du 和 Tanaka-Ishii [6]利用 Word2Vec 和 BERT 为文章的新闻标题创建双重表示，并对股票价格进行编码，再通过双向GRU处理，并使用多层感知机进行二元分类。然而，他们没有探究批量归一化和标签平滑等深度学习技术如何进一步提升泛化能力。批量归一化通过归一化步骤固定层输入的分布，从而减少模型的内部协变量偏移，并降低梯度对参数尺度或初始值的依赖 [17]。标签平滑在计算损失时，会用软目标替代独热编码（0 或 1）目标。这些软目标是原始标签与均匀分布的线性组合 [18]。Szegedy 等人 [19]首次证明，这种方法能显著提升模型的泛化能力。因此，我们希望探究这些广受欢迎且此前已证明有效的深度学习技术会如何影响 Du 等人的工作。

III. 股票价格变动分类模型

本节将说明价格变动分类问题，简要介绍如何运用自然语言处理技术生成新闻文章嵌入，并概述提出的模型。

A. 价格变动分类问题

对于第 t 天的收盘价 p_t（其中 $t \in \{1, 2, ..., T\}$），且 T 为所考虑时间段 [6], 内的交易日数，目标类别 y_t 的确定方式如下：

$$y_t = \begin{cases} 1, & p_t \geq p_{t-1} \\ 0, & p_t < p_{t-1} \end{cases} \tag{1}$$

与 Du 和 Tanaka-Ishii [6], 一样，我们将该问题视为二元分类问题。对于时间窗口 $[t-s+1, ... t-1, t]$（其中 s 为窗口大小），我们考虑第 t 天附近的时间窗口，而不是 $t-1$。将输入窗口 $[t-s+1, .., t-1]$ 输入模型以输出目标类别 t，这属于预测，因为模型不会使用第 t 天发布的文章来预测当天的价格变动。许多研究（ [1] [2], 包括 [6]）援引有效市场假说，认为这一预测问题非常困难。由于缺乏足够及时的信息来判断价格变动和文章发布哪个先发生 [6], 我们也选择研究分类问题。不过，我们还使用预测问题测试了模型，并将其与分类问题进行了比较，具体内容将在第五节说明。

B. 词与句子嵌入

为获取词级别的向量，我们使用了 Word2Vec 的 CBOW 版本 [1]，因为该版本训练速度更快，对于高频词的准确率也略高 [10]。模型使用新闻文章标题的数据进行训练；随后，对于每条标题 i，我们都使用 Word2Vec 模型生成维度为 60 的词嵌入 w_k。标题中的所有词嵌入进一步转换为关键

[1]我们使用了 gensim 库中的 Word2Vec 版本。可访问：https://radimrehurek.com/gensim/models/word2vec.html

向量 n_i^K，维度为 60，并使用词频-逆文档频率得分（TF-IDF） γ_k [6]。关键向量定义如下：

$$n_i^K = \frac{\sum_k \gamma_k w_k}{\sum_k \gamma_k}. \tag{2}$$

TF-IDF 是一种统计量，它通过计算某个词在文档中出现的次数，以及该词在文档集合中出现的文档数，来提高文档中独特词语的重要性 [20]



图 1：与感兴趣的股票相关的词簇可视化：我们根据公司所属的市场板块对公司集合进行排序，并使用 t-SNE [21] 降低向量的维度并将其可视化，从而找到与各家公司相关的相似词簇。属于同一板块的公司的向量往往分布在相近的空间中，这进一步印证了嵌入可用于发现数据的特征相似性。

公式（2）将新闻标题表示为词级嵌入。不过，它无法获取新闻标题本身的语境。例如，否定词可能会改变一个句子的含义。为了以更具语境敏感性的方式捕捉含义，我们将 BERT 编码器用作服务，以输出 1024 维向量。BERT 的嵌入输出通过 Transformer 中的自注意力机制生成 [14] [15]。随后，我们使用主成分分析（PCA） [6]，将维度降至 256，从而减少神经网络中的参数数量，并从文章的新闻标题中生成值向量 n_i^V。

随后，每篇新闻文章 n_i 都会被转换为一对向量（n_i^K，n_i^V），其中利用 Word2Vec 得到词级别的“键”向量 n_i^K，利用 BERT 得到上下文级别的“值”向量 n_i^V。为节省计算时间，这些向量只计算一次，并存储在数据集中，其中对于每一天 t，分别有一对键/值向量集合 $N_t^K = \{n_i^K\}_t$ 和 $N_t^V = \{n_i^V\}_t$。

训练好的股票嵌入由原论文的作者提供给我们 [6]。设每个股票嵌入为 s_j，其中 $j = 1, 2, ..., J$，且 $J = 50$（从 R&B 数据集中选取的股票数量）。我们利用它们计算市场向量。为了重现上一篇文章中旨在展示股票嵌入基本能力的结果，我们使用点积作为内积[6]。

这是通过文本特征提取流程 [6]完成的。首先计算每天 t股票 s_j的每篇文章 n_i 对应的注意力分数。

$$score_{i,j} = n_i^K \cdot s_j \tag{3}$$

然后，通过以下方式求得每篇文章 i 相对于股票 j 的权重：

$$\alpha_i^j = Softmax(score_{i,j}), \tag{4}$$

最后，通过以下方式计算第 t 天相对于股票 j 的市场向量 m_t^j：

$$m_t^j = \sum_{n_i^V \in N_t^V} \alpha_i^j n_i^V \tag{5}$$

针对股票 j，我们为窗口 s 中的每个交易日计算市场向量。然后，我们创建了一个向量序列 $M_{[t-s+1,t]}^j = [m_{t-s+1}^j, m_{t-s+2}^j, ..., m_{t-2}^j, m_{t-1}^j, m_t^j]$，将其作为股价变动分类器的输入，并将第 t天的目标类别设为 y_t。对于这个问题，我们发现窗口 s 的最佳取值为 5。在为期五天的序列中，正确标签在股价上涨和下跌之间的分布相当均衡（分别为 55% 和 45%）。

C. 模型概述

提出的模型以一系列市场向量作为输入，$M_{[t-5,t]}^j = [m_{t-5}^j, m_{t-4}^j, ..., m_{t-1}^j, m_t^j]$ 这些向量对应于股票 j，用于对价格变动 [6]进行分类。随后，模型输出分类值 $\hat{y}_t$。模型包含一个双向门控循环单元（Bi-GRU）层，该层接收 $M_{[t-5,t]}^j$ 和初始隐藏状态 h_0，并利用它们构建双向编码向量 h_t^O：

$$h_t^O = GRU(M_{[t-5,t]}^j, h_0) \tag{6}$$

其中，t 是我们要进行分类的交易日。GRU 的输出 h_t^O随后输入批归一化层[17]和丢弃层。经过丢弃层后，分类器通过以下方式估计概率：

$$\hat{y}_t^j = softmax(MLP(dropout(batchnorm(h_t^O)))) \tag{7}$$

其中，MLP 是用于预测二元分类 [6]的全连接层。随后，模型计算目标类别 y_t^j 与股票 j的估计值 $\hat{y}_t^j$ 之间的交叉熵损失，表示为：

$$l_j = -\frac{1}{T}\sum_{t=1}^{T}(y_t^j log(\hat{y}_t^j) + (1-y_t^j)log(1-\hat{y}_t^j)), \quad (8)$$

其中，T 是所考虑的总天数。所有股票损失 j 的均值即为损失，表示为 $l = (\sum_{j=1}^{J} l_j)/J$，其中$J$ 是所考虑的股票总数。

与 Du 等人 [6], 一样，股票走势分类面临的一个问题是，单只股票分类无法提供足够的数据来获得良好的性能。为了解决这一问题，我们在所有股票上训练一个分类器 [6]。在这种方法中，所有股票共用一个分类器。这使模型具有更强的泛化能力，能够避免样本量较小时出现的过拟合问题。

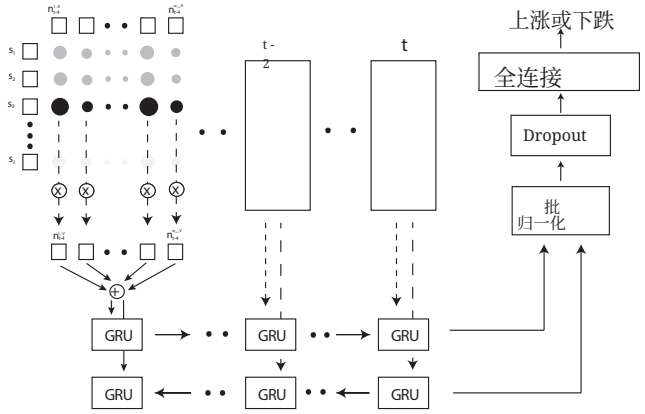

图2：所提出的深度学习模型的系统示意图，假设 $t-4$ 到 t 为交易日。不同大小的实心圆表示不同的注意力权重。

D. 训练细节

模型采用两个阶段进行训练，且学习率各不相同：探索阶段 和利用阶段。我们将这种方法称为双阶段训练。

探索阶段使用的学习率为 0.0001，利用阶段使用的学习率为 0.000001。模型在探索阶段训练了 100 个训练轮次。随后，使用验证集表现最佳的模型在利用阶段额外训练 100 个训练轮次。图3展示了一个模型在两个阶段中各训练轮次的性能示例。

训练和验证均使用包含 64 个样本的小批量。Dropout 设置为 0.2。由于这是一个将样本分类为“股价上涨”或“股价下跌”的分类问题，且输出为 Softmax，因此使用交叉熵作为损失函数。使用 Adam优化器训练模型，权重衰减设为 0.000001，学习率如上所述。

E. 标签平滑

在处理具有 Softmax 输出的监督学习问题时，标签通常以由 0 和 1 组成的向量进行独热编码，其中 1 表示正确标签。

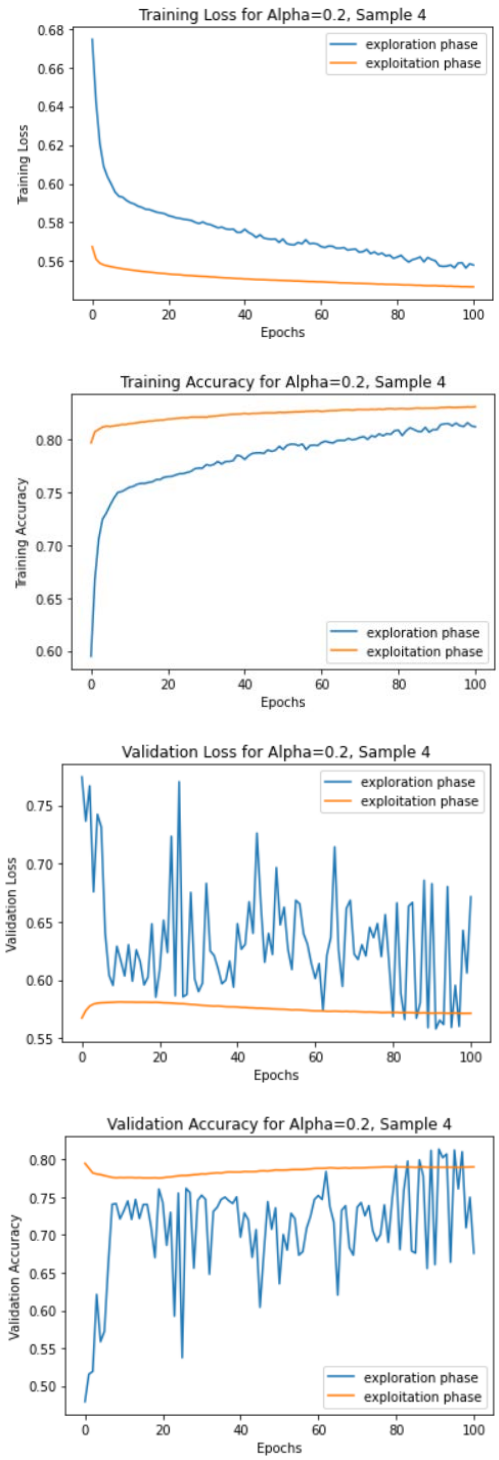

图3： α=0.2 的训练集和验证集损失与准确率示例

然后最小化输出与正确标签向量之间的交叉熵损失。

正如 Szegedy 等人 [19]最初所展示的，标签平滑正则化可显著提升模型在分类任务中的表现，此后也广受欢迎。标签平滑是一种在计算交叉熵损失 [18]时，将独热编码标签按 $y_k^{LS} = y_k(1-\alpha) + \alpha/K$针对第 k 类进行调整的技术。此处， K 是类别数量， y_k 和 y_k^{LS}分别是标签和调整后的标签，而 α 是标签平滑参数。在我们的情形中， $K = 2$ ，因为这两个类别分别表示股价明显上涨或明显下跌。

IV. 实验结果

A. 数据集

我们使用路透社与彭博社 （R&B）数据集[2] 来生成标题向量。该数据集共包含 552,909 篇文章，覆盖 2,605 天，其中有 1,794 个交易日。我们使用雅虎财经获取了 50 只股票的价格数据，这些股票在 100 篇或更多不同新闻文章中被提及 [6]。随后，我们为每个样本创建了一组交易日， $N_{[t-s+1,t]}$，其中 s是时间窗口的大小。每个样本由一个元组（$N_{[t-s+1,t]}$， y_t）构成。

原论文提出使用连续两天之间的对数收益率阈值来计算样本的模糊性 [6]。我们对每个样本中的 $[t-1, t]$ 进行了这项计算。剔除存在模糊性的样本后，样本总数为 32,204。这些样本用于获取每天每篇文章的向量，并利用 Du 和 Tanaka-Ishii 提供的股票嵌入计算市场向量 [6]。

每次训练迭代时，数据集都会随机划分为 60% 的训练集、20% 的验证集和 20% 的测试集。此操作使用 sklearn 的“train test split”函数完成，并将“shuffle”参数设为“True”。因此，每次开始新的训练迭代时，训练集、测试集和验证集中的样本都可能与之前的训练迭代不同，所包含的五日窗口也可能不同。

B. 评估指标

我们采用标准的准确率（Acc）和马修斯相关系数（MCC）作为评估指标，因为其他研究也使用这些指标来评估股票预测 [22] [23]。

准确率定义如下：

$$Acc = \frac{NumberofCorrectPredictions}{TotalNumberofPredictionsMade} \tag{9}$$

马修斯相关系数通过衡量混淆矩阵中正确分类的样本数占总样本数的比例，避免了类别分布偏斜带来的偏差。混淆矩阵包含各类样本的数量

[2]Ding 等人 [6]，中使用的数据集已公开，可从 https://github.com/WenchenLi/news-title-stock-prediction-pytorch 获取。

分类为真阳性（tp）、假阳性（fp）、真阴性（np）和假阴性（fn）的样本数。马修斯相关系数的计算公式为：

$$MCC = \frac{(tp * tn) - (fp * fn)}{\sqrt{(tp+fp)(tp+fn)(tn+fp)(tn+fn)}} \tag{10}$$

C. 分类结果

按照第 IV.A 节所述，模型评估通过对十次运行的结果取平均值完成，每次运行时均随机划分数据集。提出的模型的平均测试准确率达到 80.7%，平均 MCC 分数为 0.631。与 Du 和 Tanaka-Ishii（68.8%） [6] 以及 Ding 等人（准确率：65.1%，MCC：0.436） [23], 的表现相比，这是一次显著提升，因为我们的模型能够正确分类的样本多出 10% 以上。表 I 汇总了测试准确率、标准差和 MCC 分数。

模型：	测试准确率	σ	MCC
提出的模型	**0.807**	0.962	**0.631**
Du 和 Tanaka-Ishii [6]模型	0.688	1.67	-
Ding 等人 [23] 模型	0.651	-	0.436

表 I：测试集评估结果，对比了提出的模型与先前模型。

D. 标签平滑结果

正如 Szegedy 等人 [19] 所示，标签平滑正则化会对模型性能产生显著影响。我们的结果也印证了这一发现。表 II 展示了对独热编码标签应用不同 α时的测试集准确率。模型采用上述相同的双阶段训练方法进行训练，并使用相同的参数。进行了五次不同的运行，每次都对训练集、验证集和测试集进行打乱，并对结果取平均。

α:	0	0.05	0.1	0.2	0.3	0.4	0.5
运行 1	0.774	0.794	0.800	0.790	0.787	0.793	0.780
运行 2	0.768	0.781	0.790	0.816	0.804	0.817	0.800
运行 3	0.781	0.780	0.791	0.801	0.804	0.803	0.793
运行 4	0.742	0.774	0.794	0.791	0.798	0.790	0.786
运行 5	0.786	0.774	0.794	0.791	0.798	0.790	0.786
Avg.	0.770	0.780	0.792	**0.804**	0.798	0.803	0.789

表 II：使用不同 α 值进行标签平滑的测试集评估准确率。

对于 $\alpha \in [0.1, 0.2, ...0.4]$，标签平滑使模型在测试集上的准确率提高了 2-3%。与 Du 和 Tanaka-Ishii 达到的测试集准确率（68.8%）相比，这提高了 10% 以上 [6]。然而，随着α 趋近于 1，准确率开始下降。这是预料之中的，因为当 $\alpha \to 1$时， $y_k^{LS} \to 0.5$ 对两个类别都如此，使它们对模型而言无法区分。

E. 消融研究

我们还通过消融研究，移除若干组件，以考察它们对模型性能的影响。模型在标签平滑参数设为 0.2 时，在测试集上的表现最佳，因此采用了该 α。

采用了与上述相同的双阶段训练方法和参数。使用打乱后的训练集、验证集和测试集进行了五次重复实验，并对结果取平均。

组件 已删除	None	权重 衰减	Dropout	批次 Norm	批归一化 + Dropout
运行 1	0.790	0.781	0.794	0.745	0.741
运行 2	0.816	0.807	0.806	0.742	0.743
运行 3	0.801	0.811	0.800	0.740	0.745
运行 4	0.824	0.799	0.825	0.746	0.742
运行 5	0.791	0.802	0.800	0.729	0.746
平均值:	0.804	0.800	0.805	0.744	0.743

表 III：移除选定组件后测试集评估所得的准确率 alpha=0.2。

如表 III 所示，将权重衰减和 Dropout 都设为 0.0，对模型预测测试集股价变动的能力影响不大。事实上，移除 Dropout 反而使性能略有提升。然而，从模型中移除批归一化层，导致测试集准确率骤降约 6%。同时移除批归一化层和 Dropout 的效果，与仅移除批归一化层大致相同。

V. 讨论

A. 股票嵌入与数据的关系

这个问题有一个合理的顾虑：Du 和 Tanaka-Ishii [6] 提供的经训练的股票嵌入，使用的数据与我们训练自己模型时所用的数据相同。为了解决这一问题，我们在股票嵌入中加入了随机噪声，其尺度为嵌入标准差的五分之一；然后使用双阶段训练方法和相同的参数，基于这些新嵌入构建的数据集训练模型。噪声模型的准确率为 81.1%，MCC 分数为 0.639。

模型:	测试准确率	MCC
提出的模型	0.807	0.631
噪声模型	**0.811**	**0.639**
差异。	0.004	0.008

表 I：测试集评估结果，对比所提出的模型与噪声模型。

为了测试数据的依赖性，我们还检查了不同的窗口大小，如图4所示。结果表明，只要存在上下文，我们的系统对天数并不敏感。尽管如此，窗口大小为1时的预测效果仍优于先前的方法。由于我们使用了双向门控循环单元，反向门控循环单元不会随窗口大小而变化。不过，对于窗口大小为2或更大的情况，系统仍能学习到有效的特征。

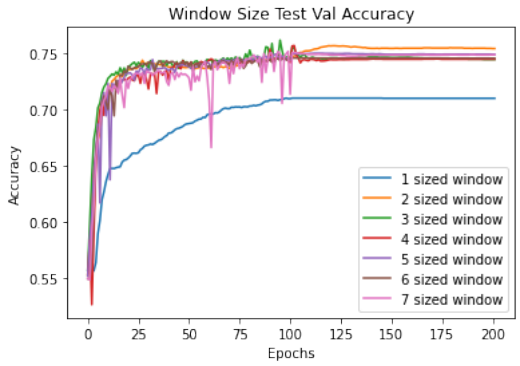

图 4：未使用批归一化时不同窗口大小下的验证准确率

B. 验证集波动

从图3可以清楚地看出，在探索阶段，当学习率较高（0.0001）时，验证集上的准确率和损失会剧烈波动。然而，如果我们查看移除批量归一化后的验证准确率和损失，就会发现它们变得更加稳定，如图5和图6所示。

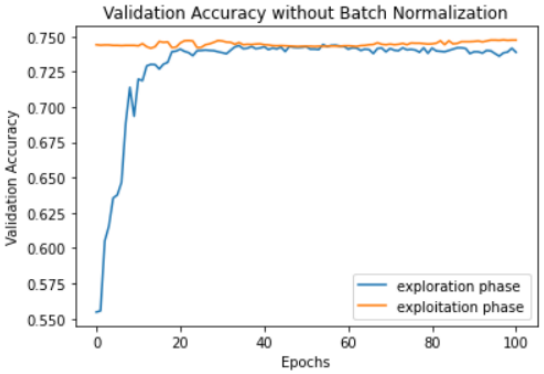

图 5： $\alpha = 0.2$ 移除批归一化后的验证准确率

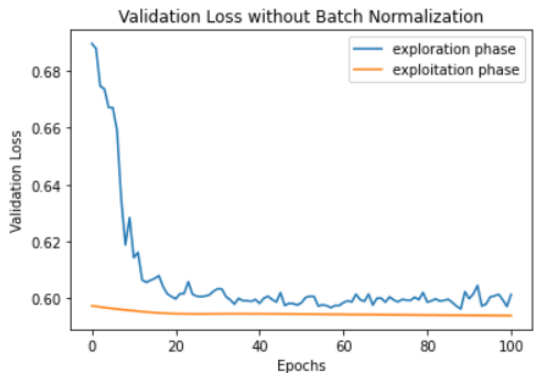

图 6： $\alpha = 0.2$ 移除批归一化后的验证损失

我们认为，这可能与批量归一化在每个训练轮次中跟踪的内部统计量有关，但还需要

进一步研究才能确认这一点。不过，表 III 清楚地表明，执行批量归一化显著提高了模型在测试集上的准确率。这表明，批量归一化确实达到了预期效果，即提升模型的泛化能力 [17]。

C. 预测与分类

为了测试模型能否预测近期的价格变动，我们使用预测数据集训练了所提出的分类模型。这涉及使用时间范围 $[t-s+1, t-1]$ 内的市场向量，并预测第 t 天的价格变动。我们将其称为预测数据集。为了证明模型能够进行预测，我们进行了如下测试：

模型：	测试准确率	MCC
使用预测集的预测模型	0.765	0.557
使用分类集的分类模型	0.817	0.644
使用分类数据集的预测模型	0.453	0.407
使用预测数据集的分类模型	0.452	0.496

表 V：比较预测与分类的测试集评估结果。

表 V 展示了我们如何在两个模型上使用预测数据集和分类数据集，以说明未为训练任务使用正确的时间范围可能导致较差的结果。分类模型无法使用预测集进行分类，反之，预测模型也无法使用分类集进行预测。图 7 展示了两个模型在此次测试训练期间的准确率。

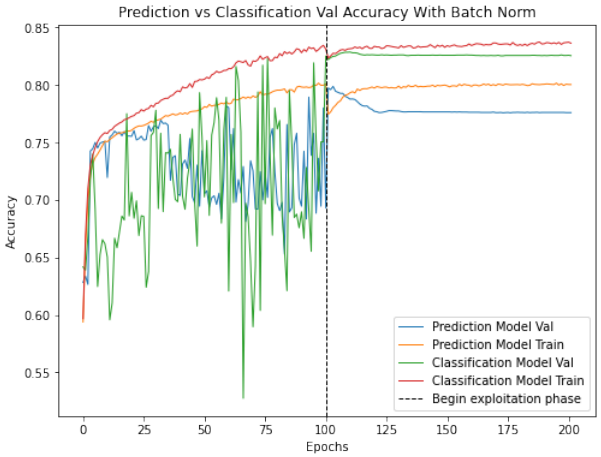

图 7：图表显示，提出的模型既能进行预测，也能进行分类

D. 有效市场假说（EMH）的启示

这项研究或许与有效市场假说（EMH）相矛盾。该假说认为，在有效市场中，价格会充分反映所有可用的相关信息 [1]。一旦发现某种模式，它带来的优势就只会持续有限时间，因为理性交易者会调整行为以应对这些新信息，而该模式也将随之消失 [2]。

然而，近年来，行为金融学的支持者对 EMH 提出了诸多批评 [24]。经济学的一项基本假设是，个人在受到激励因素影响时会理性行事。但事实显然并非总是如此，这也是行为金融学的理论基础 [25]。因此，既然新闻和其他新信息确实会影响股价，市场就可能处于所谓的“过度反应与反应不足状态”，而了解情况的投资者或许能从中获利 [26]。同样可以合理地认为，当经济力量之间的关系变得更加复杂时，例如新闻与股票之间的关系，不同的经济参与者可能会对这些模式作出不同解读，甚至完全无法识别它们。事实上，经济学家几十年来一直在研究不完全信息问题 [27], , 而完全信息的存在则是有效市场假说（EMH）等古典经济学模型的一项基础假设 [1]。因此，鉴于这一假设并不十分符合现实，如果我们的模型能够通过不同的正则化技术，比 Du 和 Tanaka-Ishii [6]揭示更多有关新闻与股票之间关系的信息，就能带来额外的金融优势。

VI. 结论

通过修改神经网络架构并引入更多泛化技术，我们显著提升了 Du 和 Tanaka-Ishii [6], 首次提出的模型的性能，使测试准确率从 68.8% 提高到 79.3%。在损失计算过程中对独热编码标签进行平滑处理，进一步将准确率提高至最高 80.4%。消融研究证实，批量归一化对模型性能有显著影响；相比之下，引入权重衰减和丢弃层的效果则不太明显。

由于我们使用了由 Du 和 Tanaka-Ishii 的模型训练得到的相同嵌入，因此，通过将股票嵌入和新闻嵌入与网络进行联合训练，或许还能进一步提高测试集准确率。

我们还预计，通过让分类问题更加贴近实际，可以提高准确率。一种简单的做法是将类别从“上涨”和“下跌”扩展为“上涨”“下跌”和“无变化”。这需要将价格变动较小的股票重新纳入数据集；这些股票最初被 Du 和 Tanaka-Ishii [6]剔除了。

此外，完全放弃独热编码，改用软标签分配 （SLA）可能会进一步提高准确率。正如 Alishahraniet al.[28], 最初所述，在此情况下，SLA 将涉及估算每只股票的股价走势分类器的概率分布。随后，在计算损失时，将为每个窗口中的每只股票分配各分类器对应的概率。

参考文献

[1] E. F. Fama, 《有效资本市场：理论与实证研究综述》, 《金融学杂志》, 第25卷，第2期，第383–417页，1970年。[在线]。网址：http://www.jstor.org/stable/2325486

[2] A. Timmermann 和 C. W. Granger，《有效市场假说与预测》，《国际预测学杂志》，第20卷，第1期，第15–27页，2004年。 [在线]。网址：https://www.sciencedirect.com/science/article/pii/S0169207003000128

[3] Q. Li、Y. Chen、J. Wang、Y. Chen 和 H. Chen，《网络媒体与股票市场：大数据视角下的调查与未来方向》，IEEE知识与数据工程汇刊，第30卷，第2期，第381–399页，2017年。[4] G. P. C. Fung、J. X. Yu 和 H. Lu，《文本信息对金融市场的预测能力》，IEEE Intell. Informatics Bull.，第5卷，第1期，第1–10页，2005年。[5] J. Bollen、H. Mao 和 X. Zeng，《Twitter情绪可预测股票市场走势》，《计算科学杂志》，第2卷，第1期，第1–8页，2011年。[6] X. Du 和 K. Tanaka-Ishii，《从新闻文章和价格历史数据中获取的股票嵌入及其在投资组合优化中的应用》，载于《计算语言学协会第58届年会论文集》。在线：计算语言学协会，2020年7月，第3353–3363页。 [在线]。网址：https://aclanthology.org/2020.acl-main.307

[7] 2020 [在线]。“斯坦福问答数据集：结果”。在线。网址：https://rajpurkar.github.io/SQuAD-explorer/ [8] T. Wilson、P. Hoffmann、S. Somasundaran、J. Kessler、J. Wiebe、Y. Choi、C. Cardie、E. Riloff 和 S. Patwardhan，《OpinionFinder：一个用于主观性分析的系统》，载于《HLT/EMNLP 2005交互式演示》，2005年，第34–35页。[9] T. Mikolov、I. Sutskever、K. Chen、G. S. Corrado 和 J. Dean，《词语和短语的分布式表示及其组合性》，载于NIPS，2013年。[10] T. Mikolov、K. Chen、G. S. Corrado 和 J. Dean，《在向量空间中高效估计词表示》，载于ICLR，2013年。[11] M. Peters、M. Neumann、M. Iyyer、M. Gardner、C. Clark、K. Lee 和 L. Zettlemoyer，《深度上下文化词表示》，载于NAACL-HLT，2018年。[12] S. Hochreiter 和 J. Schmidhuber，《长短期记忆》，载于《神经计算》，1997年。[13] K. Cho、B. van Merrienboer、C. Gulcehre、D. Bahdanau、F. Bougares、H. Schwenk 和 Y. Bengio，《使用RNN编码器-解码器学习用于统计机器翻译的短语表示》，2014年。[14] J. Devlin、M.-W. Chang、K. Lee 和 K. Toutanova，《BERT：用于语言理解的深度双向变换器预训练》，载于NAACL-HLT，2019年。[15] A. Vaswani、N. Shazeer、N. Parmar、J. Uszkoreit、L. Jones、A. N. Gomez、Ł. Kaiser 和 I. Polosukhin，《注意力机制就是一切》，载于NIPS，2017年。[16] Z. Hu、W. Liu、J. Bian、X. Liu 和 T.-Y. Liu，《聆听混沌低语：一种面向新闻的股票趋势预测深度学习框架》，2019年。[17] S. Ioffe 和 C. Szegedy，《批量归一化：通过减少内部协变量偏移加速深度网络训练》，载于《国际机器学习会议》。PMLR，2015年，第448–456页。[18] R. M¨uller、S. Kornblith 和 G. Hinton，《标签平滑何时有帮助？》arXiv预印本 arXiv:1906.02629，2019年。[19] C. Szegedy、V. Vanhoucke、S. Ioffe、J. Shlens 和 Z. Wojna，《重新思考用于计算机视觉的Inception架构》，载于《IEEE计算机视觉与模式识别会议论文集》，2016年，第2818–2826页。[20] C. D. Manning 和 H. Sch¨utze，《统计自然语言处理基础》。马萨诸塞州剑桥：麻省理工学院出版社，1999年。 [在线]。网址：http://nlp.stanford.edu/fsnlp/ [21] L. van der Maaten，《通过保留局部结构学习参数化嵌入》，载于《第十二届国际人工智能与统计会议论文集》，机器学习研究论文集系列，D. van Dyk 和 M. Welling 编，第5卷。希尔顿克利尔沃特海滩度假村，美国佛罗里达州克利尔沃特海滩：PMLR，2009年4月16日至18日，第384–391页。 [在线]。网址：https://proceedings.mlr.press/v5/maaten09a.html [22] Y. Xu 和 S. B. Cohen，《基于推文和历史价格的股票走势预测》，载于《计算语言学协会第56届年会论文集》（第1卷：长篇论文），2018年，第1970–1979页。[23] X. Ding、Y. Zhang、T. Liu 和 J. Duan，《事件驱动的股票预测深度学习》，载于《第24届国际人工智能会议论文集》，IJCAI’15系列。AAAI出版社，2015年，第2327–2333页。

[24] B. G. Malkiel，《有效市场假说及其批评者》，《经济展望杂志》，第 17 卷，第 1 期，第 59–82 页，2003 年。[25] J. R. Ritter，《行为金融学》，《太平洋盆地金融杂志》，第 11 卷，第 4 期，第 429–437 页，2003 年，选自第十四届年度 APFA/PACAP/FMA 金融会议论文集。 [在线]。网址：https://www.sciencedirect.com/science/article/pii/S0927538X03000489 N.

[26] Rezaei 和 Z. Elmi，《股价预测中的行为金融模型与行为偏差》，《数学金融与应用进展》，第 3 卷，第 4 期，第 67–82 页，2018年。[27] R. C. Merton 等，《不完全信息下资本市场均衡的简单模型》，1987 年。[28] A. Alshahrani、M. Ghaffari、K. Amirizirtol 和 X. Liu，《使用 BERT 与软标签分配预测推特消息和用户的乐观／悲观倾向》，载于*2021*年国际神经网络联合会议 *(IJCNN)*。IEEE，2021 年，第 1–8 页。