利用财务新闻预测股票价格变动，使用Word嵌入和深度神经网络

杨孟鹏和惠江
电气工程与计算机科学系，
约克大学，4700 Keele Street，Toronto，Ontario，M3J 1P3，
Canada
电子邮件：提姆@ cse.yorku.ca，HJ cse.yorku.ca @

摘要

财经新闻包含有关上市公司和市场的有用信息。在本文中，我们应用流行的词嵌入方法和深度神经网络来利用财经新闻来预测市场中的股票价格变动。实验结果表明，我们提出的方法简单但非常有效，它可以仅使用历史价格信息显着提高基准系统上标准金融数据库的库存预测准确性。

1 介绍

在过去的几年中，深度神经网络（DNN）在许多数据建模和预测任务中取得了巨大的成功，从语音识别，计算机视觉到自然语言处理。在本文中，我们有兴趣将强大的深度学习方法应用于财务数据建模，以预测股票价格变动。

传统上，神经网络被用来模拟股票价格作为预测目的的时间序列，例如（Kaastra和Boyd，1991；Adya和Collopy，1991；Chan等人，2000；Skabar 和 Cloete，2002；Zhu 等人） 。，2008）。在这些早期的工作中，由于当时可用的训练数据和计算能力有限，通常浅层神经网络被用于模拟从股票价格数据集中提取的各种类型的特征，例如历史价格，交易量等，以便预测未来的股票收益率和市场回报。

最近，在自然语言处理社区中，已经提出了许多方法来探索用于股票预测的附加信息（主要是在线文本数据），例如财经新闻（Xie等人，2013；Ding等人，2014）， twitter情绪（Si et al。，2013；Si et al。，2014），财务报告（Lee et al。，2014）。例如， （Xie et al。，2013）已经提出使用语义框架解析器来概括从句子到场景以检测特定公司（正面或负面）的角色，其中具有树核的支持向量机用作预测模型。另一方面， （Ding et al。，2014）提出使用各种词法和句法约束来提取股票预测的事件特征，他们已经将线性分类器和深度神经网络都作为预测模型进行了研究。

在本文中，我们建议使用最近的单词嵌入方法（Mikolov等，2013b；Liu等，2015；Chen等，2015）从在线金融新闻语料库中选择特征，并采用深度神经网络（DNN）根据提取的特征预测未来的库存变动。实验结果表明，财务新闻得出的特征非常有用，它们可以显着提高仅依赖于历史价格信息的基线系统的预测准确性。

2 我们的方法

在本文中，我们使用深度神经网络（DNN）作为我们的预测模型，它为我们提供了轻松组合来自不同领域的各种类型特征的优势。

NAACL-HLT 2016会议记录，第374-379页，
加利福尼亚州圣地亚哥，2016年6月12日至17日.Qc 2016计算语言学协会

预处理和标准化工作。DNN模型将从历史价格信息和在线财经新闻中提取的特征作为输入，以预测未来的股票走势（向上或向下）（Peng和Jiang，2015）。

2.1 深度神经网络

本文中使用的DNN的结构是具有许多隐藏层的传统多层感知器。由L 1隐藏的非线性层和一个输出层组成的L层DNN。输出层用于模拟每个输出目标的后验概率。在本文中，我们使用整流的线性激活函数，即f（x）= max（0，x），来计算从每个隐藏层中的激活到输出，然后将其作为输入馈送到下一层。对于输出层，我们使用softmax函数来计算两个节点之间的后验概率，代表库存和库存减少。

2.2 历史价格数据的功能

在本文中，对于目标日期的每个目标股票，我们选择前五天的收盘价并将它们连接起来形成DNN的输入特征向量：$P = (p_{t-5}, p_{t-4}, p_{t-3}, p_{t-2}, p_{t-1})$，其中t表示目标日期，$p_m$ 表示日期m的收盘价。然后，我们通过根据训练集中该股票的所有收盘价计算的均值和方差对所有价格进行标准化。此外，我们还计算五天收盘价之间的一阶和二阶差异，这些差价作为额外的特征向量附加。例如，我们计算一阶差分如下：$\Delta P = (p_{t-4}, p_{t-3}, p_{t-2}, p_{t-1}) - (p_{t-5}, p_{t-4}, p_{t-3}, p_{n-2})$。以相同的方式，通过取每个ΔP中两个相邻值之间的差来计算二阶差：$\Delta\Delta P = (\Delta P_{t-3}, \Delta P_{t-2}, \Delta P_{t-1}) - (\Delta P_{t-4}, \Delta P_{t-3}, \Delta P_{t-2})$。最后，对于特定日期的每个目标库存，表示历史价格信息的特征向量由P，ΔP和ΔΔP组成。

2.3 财经新闻特写

为了从财经新闻语料库中提取适合DNN的固定大小的特征，我们需要预先
处理文本数据。对于所有金融文章，我们
首先将它们分成句子。我们只保留那些至少提到股票名称或上市公司的句子。每个句子都标有原始文章的出版日期和提到的股票名称。有可能在一个句子中提到多个股票。在这种情况下，对于每个提到的股票，该句子被标记几次。

然后，我们将这些句子按出版日期和基础股票名称分组，以形成样本。每个样本都包含在同一天发布并提及相同股票或公司的句子列表。此外，根据CRSP金融数据库（Booth，2012）咨询的第二天收盘价，每个样本被标记为正（“涨价”）或负面（“降价”）。在下文中，我们介绍了从每个样本中提取三种类型特征的方法。

(1) 一袋关键字（BoK）：我们首先根据最近的词嵌入方法选择关键词（Mikolov等，2013a; Mikolov等，2013b）。使用流行的word2vec方法 [1]，我们首先计算训练集中出现的所有单词的向量表示。其次，我们手动选择一小组种子词，即这项工作中的9个种子词，包括浪涌，上升，收缩，跳跃，跌落，跌落，暴跌，收益，暴跌，这些都被认为对股票有很强的指示作用。价格走势。接下来，我们将重复迭代搜索过程以收集其他有用的关键字。在每次迭代中，我们计算训练集中出现的其他单词与每个种子单词之间的余弦距离。余弦距离表示单词向量空间中两个单词之间的相似性。例如，基于预先计算的单词向量，我们发现了其他单词，例如反弹，下降，翻滚，减速，爬升，这些单词非常接近上述种子单词中的至少一个。选择前10个最相似的单词，并在每次迭代结束时将其添加回种子单词集。更新的种子单词将用于再次重复搜索过程以找到另外十个最相似的单词，种子单词的大小将随着我们重复该过程而增加。通过这种方式，我们搜索了训练集中出现的所有单词，最后选择了1000个单词（包括9个初始种子单词）作为我们预测的关键词

[1] https://code.google.com/p/word2vec/

任务。在这个收集关键词的迭代过程中，我们发现派生关键词的最终集合通常非常相似，只要我们从一小部分种子词开始，这些种子词都强烈表明股票价格变动。

最后，为每个样本生成一个1000维特征向量，称为关键字包或BoK。BoK向量的每个维度是针对来自整个训练语料库的每个所选关键词计算的TFIDF得分。

(2) **极**性分数（PS）：我们进一步计算所谓的极性分数（Turney和Littman，2003；Turney和Pantel，2010），以衡量每个关键词与股票变动的关系以及每个关键词如何应用于每个句子中的目标股票。为此，我们首先计算每个关键字w的逐点互信息：

$$PMI(w, pos) = log \frac{freq\ (w,\ pos)\ xn}{freq\ (w)\ x\ freq\ (pos)},$$

其中freq（w，pos）表示在所有正样本中出现的关键字w的频率，N表示训练集中的样本总数，freq（w）表示在整个训练集中出现的关键字w的总数和freq（pos）表示训练集中的阳性样本总数。此外，我们计算每个关键字w的极性分数为：

$$PS\ (w) = PMI\ (w,\ pos)\ -\ PMI\ (w,\ neg)。$$

显然，上述极性分数PS（w）衡量每个关键词与股票变动（正面或负面）的关系以及多少。

接下来，对于所有样本中的每个句子，我们需要检测每个关键字与所述股票的关系。为此，我们使用Stanford解析器（Marneffe等，2006）来检测目标股票是否是关键字的主题。如果目标股票是关键字的直接对象，我们假设关键字与基础股票相反。因此，我们需要翻转极性分数的符号。否则，如果目标股票是关键字的主题，我们保持关键字的极性分数不变。例如，在“Forrester Research 2013年客户体验调查”中，“苹果在三星和微软背后落后”这句话中包含一个已确定的关键字单据。根据解析结果，我们知道Apple是滑动的主题，而三星和微软是滑动的对象。因此，如果该句子用作Apple的样本，则直接使用上述滑动的极性分数。但是，如果将此句用作三星或微软的样本，则通过乘以1来翻转滑动的极性分数。最后，将得到的极性分数乘以TFIDF分数，以便为每个样本生成另一个1000维特征向量。

(3) **类**别标签（CT）：在财务新闻数据的预处理过程中，我们发现财务新闻中经常会描述某些类型的事件，并且在发布此类财经新闻后股票价格会发生显着变化。为了发现这些特定事件对股票价格的影响，我们进一步定义了一个类别列表，这些类别可以指示上市公司的特定事件或活动，我们称之为类别标签。在本文中，定义的类别标签包括：新产品，收购，价格上涨，降价，诉讼，财务报告，投资，破产，政府，分析师 - 亮点。首先手动为每个类别分配一些与该类别密切相关的单词。例如，我们选择发布，发布，展示，揭示新产品类别的种子词列表，这表明该公司宣布新产品。类似地，我们使用上述单词嵌入模型通过搜索与所选种子单词具有更接近余弦距离的更多单词来自动扩展单词列表。最后，我们选择前100个单词分配给每个类别标签。

在我们收集了每个类别的所有关键词之后，对于每个样本，我们计算每个类别下所有单词的出现总数，然后我们采用对数来获得特征向量为$V = (\log N_1,\ \log N_2, \log N_3, \ldots, \log N_c)$，其中
N_c 表示单词中的总次数
类别c出现在样本中。在N_c 为零的情况下，它被大的负数替换，例如在该工作中为-99.0。

2.4 通过相关图预测看不见的股票

市场上有大量股票交易。但是，我们通常只能找到一个

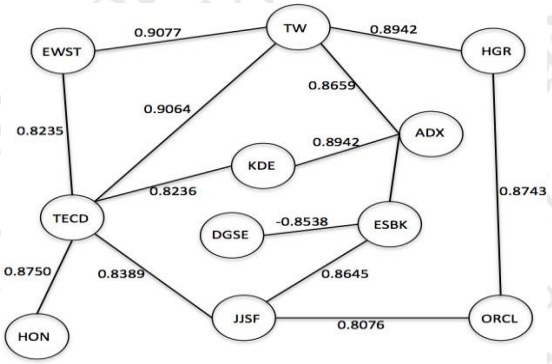

图1：相关图的一部分的图示

他们在每日财经新闻中提到的部分。因此，对于每个日期，上述方法只能预测新闻中提到的那些股票。在本节中，我们提出了一种新的方法来扩展，以预测更多可能不会在财经新闻中直接提及的股票。在这里，我们建议使用图1所示的股票相关图来预测那些看不见的股票。股票相关图是一个无向图，其中每个节点代表一个股票，两个节点之间的弧代表这两个股票之间的相关性。这样，如果某一天的新闻中提到了图表中的某些股票，我们首先使用上述方法来预测这些提到的股票。之后，预测沿着图中的弧传播，以产生对那些看不见的股票的预测。

(1) 构建图表：我们从CRSP数据库（Booth，2012）中选择前5,000个股票来构建相关图。每次选择集合中的任何两只股票以根据相关日期（2006/01/01 - 2012/12/31之间）调整其收盘价格，我们只保留具有重叠交易期的股票对至少252天（一年中的交易日数）。然后我们计算这两只股票收盘价之间的相关系数。计算出的相关系数（1和1之间）

1）附着在连接这两种股票的弧上

在图中，表明它们的价格相关性。计算来自5,000种股票的所有股票对的相关系数。在本文中，我们仅保持弧的绝对相关值大于0.8，所有其他边被认为是不可靠的并从图中修剪。

(2) 预测看不见的股票：为了预测看不见的股票的价格走势，我们首先从DNN输出中获取那些提到的股票的预测结果，通过它我们构建一个5000维向量x。设置此向量中每个维度的值以指示价格向上或向下移动的概率。对于与财经新闻中提到的股票相对应的维度，我们使用DNN的预测输出设置值。由于DNN有两个输出，每个输出代表两个类别的概率，即股票价格上涨或下跌。如果样本被识别为价格上涨，我们将此维度设置为其概率。否则，如果将其识别为降价，我们将此维度设置为其概率乘以1.0。接下来，我们为与看不见的股票相对应的所有其他维度设置零。图传播过程可以在数学上表示为矩阵乘法：

$$(X = AX^{i}$$

其中A是对称的5000乘5000矩阵，表示图中的所有弧相关权重。当然，该图传播可以重复若干次，直到预测x^{i} 收敛。

3 数据集

我们在本文中使用的财经新闻数据由（Ding et al。，2014）提供，其中包含来自路透社的106,521篇文章和来自彭博社的447,145篇文章。新闻文章发布于2006年10月至2013年12月期间。历史股票安全数据来自安全价格研究中心（CRSP）数据库（Booth，2012），该数据库由芝加哥商学院出版，被广泛用于金融建模。CRSP数据库针对所有特殊价格事件进行了适当调整，例如股票拆分和股息收益率。我们仅使用2006年至2013年的安全数据来匹配财经新闻的时间段。根据样本的发布日期，我们将数据集分为三组：训练集（2006-10-01和2012-12-31之间的所有样本），验证集（2013-01-01和2013-

06-15）和测试集（2013-06-16至2013-12-31）。训练集包含65,646个样本，验证集包含10,941个样本，测试集包含9,911个样本。

4 实验

4.1 使用DNN的股票预测

在第一组实验中，我们使用DNN来预测基于各种特征的股票价格变动，即产生第二天价格变动的极值预测（价格上涨或价格下跌）。在这里，我们使用不同的特征向量组合训练了一组DNN，发现4个隐藏层的DNN结构（每层有1024个隐藏节点）在验证集中产生最佳性能。我们仅使用历史价格功能来创建基线，并在其上添加源自财经新闻的各种功能。我们通过计算测试集上的错误率来测量最终性能。如表1所示，来自财经新闻的特征可以显着提高预测准确性，并且通过使用2.2和2.3节中讨论的所有特征，我们获得了最佳性能（错误率为43.13%）。我们还将（Ding et al。，2014）中提出的结构化事件特征应用于我们的样本，结果也列在表1中，表明我们提出的特征在预测单个股票价格池方面产生了更好的性能。

功能组合	*错误率*
价钱	48.12%
价格+博克	46.02%
价格+ BoK + PS	43.96%
价格+ BoK + CT	45.86%
价格+ PS	45.00%
价格+ CT	46.10%
价格+ PS + CT	46.03%
价格+ BoK + PS + CT	**43.13%**
结构化事件（Ding et al。，2014）	44.79%

表1：测试集上的库存预测错误率。

4.2 通过相关性预测看不见的股票

在这里，我们根据测试集上所有样本的日期对DNN的所有输出进行分组。对于每个日期，我们根据所有未观察到的股票的所有观察到的股票和零的DNN预测结果创建向量x，如2.4节所述。然后，通过相关图传播矢量以生成另一组库存移动预测。杜尔

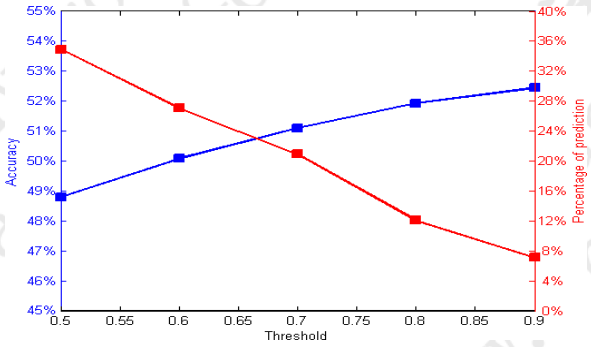

图2：通过相关性预测看不见的股票

通过传播，我们通过将向量与相关矩阵（$x^i = Ax$）相乘来计算来自不同迭代的结果。我们的实验结果表明，预测精度在第4次迭代后停止增加。在传播收敛之后，我们可以对传播的向量应用阈值以修剪所有低置信度预测。其余的可用于预测测试集上看不见的一些股票。将所有看不见的股票的预测与第二天的实际股票走势进行比较。实验结果如图2所示，其中左侧y轴表示预测精度，右侧y轴表示在各种修剪阈值下每天5000次预测的种群百分比。例如，使用较大的阈值（0.9），我们可以预测每天354个额外看不见的股票的准确率为52.44%，此外还预测测试集中每天只有110个股票。

5 结论

在本文中，我们提出了一种简单的方法来利用财经新闻来预测基于流行的单词嵌入和深度学习技术的股票移动。我们的实验表明，财务新闻在股票预测中非常有用，并且所提出的方法可以提高标准金融数据集的预测准确性。

承认

这项工作得到了Discovery Grant和加拿大自然科学与工程研究委员会（NSERC）的Engage Grant的部分支持。

参考

Monica Adya和Fred Collopy。1991.神经网络在预测和预测方面的效果如何？审查和评估。Journal of Forecasting，17：481-495。

芝加哥摊位。CRSP美国股票数据库和CRSP美国指数数据库的CRSP数据描述指南。芝加哥大学商学院安全价格研究中心。

Man-Chung Chan，Chi-Cheong Wong和Chi-Chung Lam。利用共轭梯度学习算法和多元线性回归权重初始化的神经网络预测金融时间序列。经济与金融计算，61。

陈志刚，魏琳，陈倩，陈小平，司伟，朱晓丹，惠江。2015.重新审视单词嵌入以形成对比的意义。在计算语言学协会（ACL）第53届年会的会议记录中。计算语言学协会。

小丁，张悦，刘挺，段俊文。使用结构化事件预测股票价格变动：实证调查。在2014年自然语言处理经验方法会议论文集（EMNLP），第1415-1425页。计算语言学协会。

Iebeling Kaastra和Milton Boyd。1991.设计一个神经网络，用于预测财务和经济时间序列。Neurocomputing，10：215-236。

Heeyoung Lee，Mihai Surdeanu，Bill Maccartney和Dan Jurafsky。关于文本分析对股票价格预测的重要性。在第九届国际语言资源与评估会议（LREC）会议记录中。

刘泉，惠江，司伟，凌振华，余虎。2015.学习基于序数知识约束的语义词嵌入。在计算语言学协会（ACL）第53届年会的会议记录中。计算语言学协会。

Marie-Catherine Marneffe，Bill MacCartney和Christopher D. Manning。生成类型依赖关系从短语结构解析中解析。在LREC的会议录中。

Tomas Mikolov，Kai Chen，Greg Corrado和Jeffrey Dean。2013a。向量空间中词表示的有效估计。在ICLR研讨会论文集中。

Tomas Mikolov，Ilya Sutskever，Kai Chen，Greg S Corrado和Jeff Dean。2013b。单词和短语的分布式表示及其组合性。在NIPS的会议录，第3111-3119页。

杨孟鹏和惠江。利用词汇嵌入和深度神经网络，利用财经新闻预测股票价格变动。在arXiv：1506.07220。

Jianfeng Si，Arjun Mukherjee，Bing Bing，Qing Li，Huayi Li和Deng Xiaotie Deng。2013.利用基于主题的推特情绪进行股票预测。在计算语言学协会（ACL）第51届年会的会议记录中，第24-29页。计算语言学协会。

Jianfeng Si，Arjun Mukherjee，Bing Bing，Sinno Jialin Pan，Qing Li和Huayi Li。2014.利用社会关系和情绪进行股票预测。在2014年自然语言处理经验方法会议论文集（EMNLP），第1139-1145页。计算语言学协会。

Andrew Skabar和Ian Cloete。2002.神经网络，金融交易和有效市场假说。在Proc。第二十五届澳大利亚计算机科学大会（ACSC）。

Peter D. Turney和Michael L. Littman。测量赞美和批评：从关联中推断语义取向。ACM Trans。天道酬勤。Syst。，21（4）：315-346。

Peter D. Turney和Patrick Pantel。从频率到意义：向量空间模型的语义。Journal of Artificial Intelligence Research，37（1）：141-188。

Boyi Xie，Rebecca Passonneau，Leon Wu和Germa'n G. Creamer。2013.预测股票价格变动的语义框架。在计算语言学协会（ACL）第51届年会的会议记录中，第873-883页。计算语言学协会。

朱晓天，王宏，李旭，李怀祖。通过神经网络预测股票指数增量：交易量在不同视野下的作用。Expert Systems with Applications，34：3043-3054。