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

基于预训练语言模型的股市变动预测的分析

李金阳 [1]，金明哲 [2,3*]，宿久洋 [4]

(1. Graduate School of Culture and Information Science, Doshisha University, Kyoto, Japan)
(2. Research Center for Linguistic Ecology, Doshisha University, Kyoto, Japan)
(3. Institute for General Research, Kyoto University of Advanced Science, Kyoto, Japan)
(4. Department of Culture and Information Science, Doshisha University, Kyoto, Japan)

摘要：预测股市变动是很困难的，而通过与股市息息相关的金融新闻来预测股市变动被认为是一种有效途径的同时也被认为是极具挑战性的自然语言处理任务。在股市预测任务中，新闻语料往往具有特征少噪声多的特征，用于分析这些语料的自然语言处理模型通常要具有很强的特征抽取能力的同时还能够处理超长文本。现有主流的自然语言处理模式普遍基于预训练语言模型，但预训练语言模型并不善于处理超长文本处理问题。在处理超长文本的新闻语料分析任务中，拥有更强特征抽取能力的预训练语言模型是否仍旧能够优于传统的不受文本长度限制的统计模型？本文通过搭建几种常见的文本处理模式进行实验对比，并提出了用于改善预训练语言模型的超长文本处理问题的文本提取算法。经过实验，现有的预训练语言模型相对于传统的统计模型，在处理超

基金项目：日本文部省科研基金(研究课题号 22K12726）

【作者简介】

李金阳：研究方向为机器学习和自然语言处理，ctmg0005@mail4.doshisha.ac.jp
金明哲：研究方向为文本挖掘和数据挖掘，mjin@mail.doshisha.ac.jp
宿久洋, 研究方向为统计科学和机器学习，hyadohis@mail.doshisha.ac.jp
*通讯作者：金明哲

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

长文本的新闻语料预测股价变动的任务上并没有显著的优势，而本文提出的文本平均提取算法则能够提高预训练语言模型约 3%的准确度。
关键词：金融新闻, 股市预测, 预训练语言模型, 长文本处理, 文本提取

Analysis of Stock Market Movement Prediction with Pre-trained Language Model

Jinyang Li[1]，Minzhe Jin[2,3]，Yadohisa Hiroshi[4]

(1. Graduate School of Culture and Information Science, Doshisha University, Kyoto, Japan)
(2. Research Center for Linguistic Ecology, Doshisha University, Kyoto, Japan)
(3. Institute for General Research, Kyoto University of Advanced Science, Kyoto, Japan)
(4. Department of Culture and Information Science, Doshisha University, Kyoto, Japan)

Abstract: Predicting stock movements is undeniably difficult, and since financial news is so closely related to the stock market, it is considered an effective way to predict stock movement while a challenging NLP task. Because financial news corpus always comes with very few features and lots of noise, the models must be capable of being strong in feature extraction and handling ultra-long texts. The mainstream natural language processing patterns are generally based on pre-trained language models (PLM), but PLMs are not good at processing ultra-long texts. So, will the PLMs outperform the traditional statistical models in handling ultra-long news corpus? In this paper, several typical NLP patterns are built for experimental comparison, and a text extraction algorithm aiming to improve the ultra-long text handling problem is proposed. According to the result, the PLMs have no significant advantage in analyzing ultra-long news corpus compared with traditional statistical models, and the text extraction algorithm we proposed can improve the accuracy of the pre-trained language model by about 3%.
Keywords: Financial news, Stock market prediction, Pre-trained language model, Ultra-long text processing, Text extraction

1 引言

众所周知，股票市场涉及因素多，形势变动速度快，信息量大处理复杂，被称为金融学领域最难预测的内容之一。股价的变动往往被称为事件驱动，因为受市场发生的大大小小的事件都有可能导致股价的波动。但是现实世界中准确预测事件的发生是极其困难的，而事件发生所产生的新闻则为预测这些股价变动提供了一个更简单的途径。新闻普遍以文本为载体，新闻文本语料来预测股价变动就被认为是一种有效的手段。

目前的文本分析技术发展快速，很多研究者都采用了最新的技术来提高预测的精准度，但对于特征量少而噪声多的新闻语料来说，现有的研究都是额外增加特征量，并没有针对噪声多这个问题做出多少有效的改善。此外，最新的文本分析技术，如大部分预训练语言模型，虽然特征提取能力更强但并不擅长于处理超长文本，在需要处理超长文本的新闻语料分析任务中，这些新的文本分析技术是否优于传统的分析技术，程度如何，仍然是一个需要被研究的话题。

本文针对先行研究这两个问题，搭建了基于词频特征，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进行对比实验，并提出了用于改善预训练语言模型的超长文本处理问题的文本提取算法。实验结果表明，现有的预训练语言模型相对于传统的统计模型，在处理新闻语料预测股价变动的任务上并没有显著的优势。本文提出的文本提取算法则能够平均提高预训练语言模型约 3%的准确度。

2 相关工作

整体而言，近年来自然语言处理技术有了明显的发展。早期的文本分析任务中特征提取通常基于词袋(Bag-of-words)，TF-IDF 值[1]，n-gram 模型，词典等方法。Mikolov 等[2]在 2013 年提出了 CBOW 和 Skip-gram 模型，通过训练这两种模型可以高效地得到能够包含上下文信息的静态词向量表现，为后续研究提供了“词的定义是由上下文决定的”的研究思路，Vaswani 等[3]在 2017 年提出了 Transformer 模型，这种带有注意力机制的模型所包含的 Encoder 和 Decoder 结构能够高效的抽取和分析特征。Peters 等[4]在 2018 年提出了使用 bi-LSTM 模型及预训练模式产生动态词向量的方法，这种动态词向量能够解决静态词向量难以处理的多义词问题，同年 Devlin 等[5]提出了 BERT 模型，同样采用预训练模式的同时，所使用的带有自注意力机制的 Encoder 具有更强的特征抽取能力的同时还能够有效避免 RNN 类模型的长期依赖问题。在 BERT 的基础上，2019 年 Liu 等[6]进一步改进预训练模式和输入方式，提出了 RoBERTa 模型进一步提升了模型性能，同年 Araci[7]通过进一步在金融语料上进行预训练，发布了 FinBERT 模型专门用于金融领域的自然语言处理任务。

随着自然语言处理技术的发展，使用新闻等文本语料预测股市的研究也在同步进行。Schumaker 等[8]在 2009 年使用了词袋等文本表示并使用 SVM 分类器分析了新闻语料并预测了股价变动方向，Ding 等[9]使用结构化表示来处理新闻语料并构建卷积神经网络进行预测，Pagolu 等[10]使用 CBOW 模型构建了静态词向量表示，Hu 等[11]在 2017 年第一次将 Transformer 模型代入股价预测任务中，Zhou 等[12]在 2020 年提出了 POT 机制并使用 BERT 模型对股市进行预测，还发布了一段新闻语料，Chen[13]在 2021 年提出了一种可以产生用于股市预测动态词嵌入的基于 BERT 的 FT-CE-RNN 模型，Lin 等[14]在 2022 年研究了影响股价预测的三个因素，包括文本表现方式，分类器和信息源，并发现其对预测的影响并不小，Li 等[15]则提出了一种由全连接层连接的包含两个采用不同方法训练的循环神经网络的混合模型来预测股价，Villamil 等[16]在 2023 年提出一种基于 biLSTM 来获取文章和股价关系的算法，以此来提高模型的泛化性。

自然语言处理技术的进步，虽然在文本分析领域内大多数任务都取得了巨大的进步，但是在需要处理以噪声多为特点的新闻语料中，现有的研究通常都是以增加外部信息或者特征量来提高精度，这对于噪声过多的问题并没有改善。而且现有的研究当中，不同的研究所采取语料不同，预测对象不同，评价标准不同[17]，不同的研究之间的比较很困难，在需要处理超长文本新闻语料的股市预测任务中，并不擅长于处理超长文本的预训练语言模型是否真的好过没有文本长度限制的统计模型，这仍然是一个需要被研究的内容。

3 方法及原理

3.1 文本分类模式

现有的文本分类模式根据所使用的文本特征量的不同，可分为基于词频特征的分类模式，基于静态词嵌入的分类模式和基于动态词嵌入的分类模式。这些分类模式分别发展于不同的时期，具体而言有多种实现方式，本文选取不同模式中较为有代表性的实现。

3.1.1 基于 Bag-of-Words 特征量的分类模式

基于词频特征的分类模式指的是使用基于词频的简单算法来提取文本的特征并将其转化为向量，再使用分类器来得出分类结果的文本分类模式。

Bag-of-Words（BOW）是一种最简单的词频表示方法，将一段文本看作一个词的集合，忽略词与词之间的关系，仅仅统计每一个词的出现次数。

针对每篇文本，BOW 会生成一个长度为语料总单词数的固定向量，因此 BOW 所搭配的分类器可以是任意的可以处理固定长度向量的分类器。本文使用了 Linear Classifier（LM），Random Forest（RF）和 Support Vector Machine（SVM）作为分类器。

BOW 不受到文本长度的限制，但仅仅是词频的信息，文本中很重要的上下文信息则被完全丢弃，实际利用局限性很大。

3.1.2 基于 TF-IDF 特征量的分类模式

Term frequency–inverse document frequency(TF-IDF)算法[1]是一种典型的基于词频的特征提取算法。

TF-IDF 是一种常见的统计量，主要用于评估一个词对于一个文本集或一个语料库中的其中一个文本的重要程度，对于每一个单词i在文本j中其 TF-IDF 值定义为

$$TFIDF_{i,j} = TF_{i,j} \times IDF_i \tag{1}$$

其中，$TF_{i,j}$定义为

$$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{2}$$

其中，$n_{i,j}$表示单词i在文本j中出现的次数，$\sum_k n_{k,j}$表示文本 j 中出现的单词总数。

IDF_i的定义为

$$IDF_i = \lg \frac{D}{d_i} \tag{3}$$

其中D表示语料库中文本总数，d_i表示单词i出现过的文本总数。

由定义可知，TF-IDF 值和单词在文本中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降，因此 TF-IDF 值更加倾向于抽取出在某些文本中单独多次出现的单词，可以很好的抽取出文本中独特的代表词。

针对每篇文本，TF-IDF 会生成一个长度为语料库总单词量的固定向量，因此 TF-IDF 所搭配的分

类器可以是任意的可以处理固定长度向量的分类器。本文使用了 LM，RF 和 SVM 作为分类器。

TF-IDF 算法简单高效且易于实现，和 BOW 类似，作为一种纯粹基于词频特征的提取算法，几乎不受到文本长度的限制。但是同时也是因为其纯粹基于词频的特性，极其容易受到高频词的影响，而且因为仅仅含有词频的信息，文本中很重要的上下文信息则被完全丢弃，也不能算作一种好的文本特征提取算法。

3.1.3 基于 word2vec 特征量的分类模式

为了更高效地提取上下文信息，基于 “词的定义由周围的词决定”这一理念，Mikolov 等[2]提出了 Skip-gram 和 Continuous bag-of-words（CBOW）模型，基于这两种模型的实现 word2vec 就是一种有名的静态词嵌入模型。

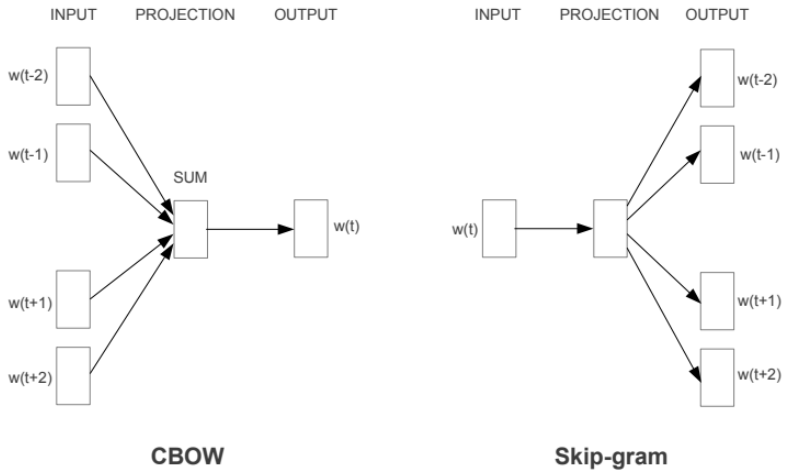

图 3. 1 Continuous bag-of-words（CBOW）和 Skip-gram 模型

其中 CBOW 模型使用周围的词来预测词，而 Skip-gram 模型使用词来预测周围的词。

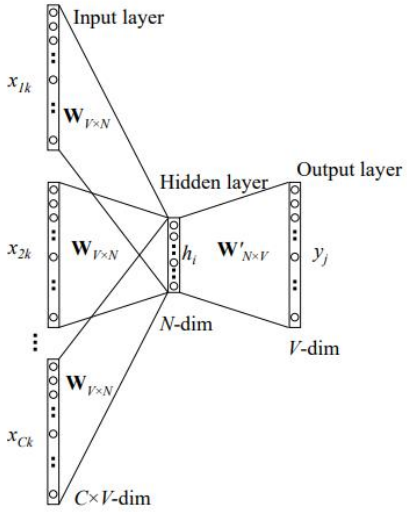

图 3. 2 CBOW 模型

如 Rong[18]所给出的图 3.2 一样，对于输出y的每一维u_j都有

$$u_j = W' \cdot h \tag{4}$$

其中W'为一个维度为$N \times V$的权重矩阵，N为隐藏层节点数，V为输入输出维度。最终u_j经过 Softmax 获得输出y。而h则为

$$h = \frac{1}{C} W(x_{1k} + x_{2k} + \cdots + x_{Ck}) \tag{5}$$

其中W为一个维度为$V \times N$的权重矩阵，$x_{1k}, x_{2k}, \ldots, x_{Ck}$为输入，$C$为输入的数量，所有输入共享一个权重矩阵$W$。经过充分训练的$W$和$W'$即为词嵌入。

Skip-gram 模型和 CBOW 模型略有不同，但是总体而言 Skip-gram 模型可以看作 CBOW 模型的逆过程，因此此处不再赘述。

由 word2vec 所产生的词嵌入向量能够保证在相对有限的维度内保存较多的上下文信息，由此也在真正意义上保留了原词的词义。根据“词的定义由周围的词决定”原理，越相近的词其周围的词也越相近，自然其产生的词向量也越接近。根据词向量是否相似，由词向量来判断近义词与反义词成为可能。

针对每篇文本，word2vec 会生成一个维度为词向量维度*文本词数的向量序列，因此 Word2vec 所搭配的分类器一般为可以处理序列的分类器。本文使用了 RNN 和 LSTM 作为分类器。

Word2vec 的提出，带来了一种处理高效且效果较好的文本特征提取方式，真正意义上做到了将每个词的词义保留到抽取的特征当中。但是，作为一种静态词嵌入算法，word2vec 经过训练所得到的词表一经得出就不再产生变动，处理多义词等情况时，因同一词在词表对应的向量是同一个，所得到的向量也只有同一个，这就导致了不同的词义向量却是同一个的问题。且基于 word2vec 特征量的分类模式受文本的长度限制，过长的文本会导致循环神经网络的遗忘问题。

3.1.4 基于 BERT 特征量的分类模式

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[5]，在此基础上还有很多改进后的实现，也可以称为 BERT-based 模型。这些都是典型的动态词嵌入模型，且这类模型都基于预训练-微调模式进行训练和使用，也被称为预训练模型。

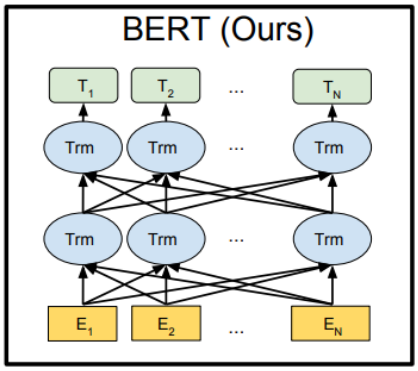

图 3.3 BERT 模型

对于每一个输入的 token 静态嵌入X_i，经过 Segment Embeddings 和 Position Embeddings 处理后得到E_i并进入多层 Encoder 并最终输出 token 的嵌入T_i。其中每一层 Encoder 的结构如图 3. 4 所示。

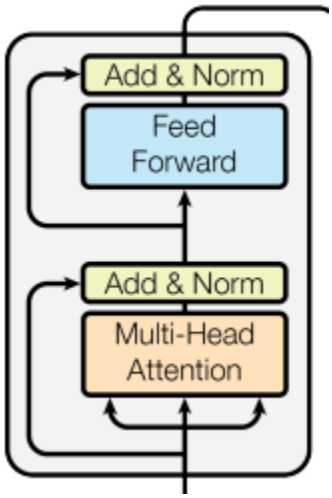

图 3. 4 Encoder 模型

具体到 Encoder 内部，由维度 d 的向量 e 组成的输入$E = (e_1, e_2, \dots, e_n)^T$的都进行 Multi-Head self-attention 操作得到$E_{Attention}$

$$E_{Attention} = Softmax\left(\frac{Q \times K^t}{\sqrt{d}}\right)V \tag{6}$$

其中Q, K, V由输入向量和三个维度为$Multi-Head$数$\times$ 最大文本长度$\times$ $WordEmbedding$维数的权重矩阵相乘得来。

$$\begin{aligned} Q &= W_Q E \\ K &= W_K E \\ V &= W_V E \end{aligned} \tag{7}$$

其中Q, K, V均为$Multi-Head$数$\times$ 最大文本长度$\times$ 最大文本长度矩阵。

之后将$E_{Attention}$与E连接并进行 Layer Normalization 处理。

$$E_{Attention} = LayerNorm(E + E_{Attention}) \tag{8}$$

Layer Normalization 目的是将隐藏层归一为标准正态分布

$$LayerNorm(x) = \frac{x_{i,j} - \mu_j}{\sqrt{\sigma_j^2 + \epsilon}} \tag{9}$$

其中μ_j表示矩阵j列的均值，σ_j^2表示j列的方差，ϵ为避免分母为 0 的常数。

之后$E_{Attention}$代入到 FeedForward 网络中

$$E_{hindden} = Linear(ReLU(Linear(E_{Attention}))) \tag{10}$$

最后用$E_{Attention}$连接$E_{hindden}$进行 Layer Normalization 处理得到本层 Encoder 的输出E_O

$$E_O = LayerNorm(E_{Attention} + E_{hindden}) \tag{11}$$

BERT-based模型因为其采用了来自于Transformer的Encoder模块，具有极强的特征抽取能力，同时其根据实时动态输出的模式，上下文不同所得到的词向量不同，解决了多义词同向量的问题。而Encoder模块所带有的self-attention机制，会对所有的上下文求值，使BERT-based模型能够从根本上克服各种RNN模型的长文本遗忘的问题。

根据Devlin等[5]的观点，BERT-base类的模型并不仅仅是作为一种词嵌入使用，实际上它可以负担特征抽取和分析两部分功能，因此本文根据其推荐的方法，后续接一层简单的全连接层。

BERT-based模型基于预训练-微调模式，根据其预训练所使用的语料和训练模式不同，也会影响模型的效果。

BERT-based模型拥有极佳的性能同时也具有极广泛的适用性，但是对于金融新闻分析任务来说，分析每天数十篇新闻文章，数百句中包含的上万个token这样的超长文本分类任务，极其消耗计算资源的BERT-based模型几乎不太可能完整分析。

3.2 文本提取算法

所以在使用BERT这类模型的时后，通常需要对超长文本进行处理，常用的方法是分篇或者分片段。而对于噪声量多的超长文本来，分篇或者片段处理有可能导致被分到的整部分都是噪声的情况，进而导致模型完全无法收敛的问题。解决这一问题的最好办法就是更准确的标注这些片段，但是如何大量又准确地标注这些片段又是一个问题。

而且现有的类似研究当中，通常都是增加从其他外部信息量中抽取的特征来提高精度，但是这并没有解决原本语料中噪声过多的问题。既然新闻语聊是一种特征量少而噪声多的语料。那么能不能通过一种方式，抽取出其中还有特征的句子，保留有用信息的同时，还降低文本的长度使其更容易被处理。

因此，本文基于以下想法提出一种算法。

1. 文本的特征以句子为单位。

2. 句子的特征值由句子所含的词的特征决定。

3. 词的特征值由两个因素决定，其一是词在股价上涨日出现次数和股价下跌日出现次数的差值，其二是词是不是更倾向于在股价变动激烈日出现。

具体而言，本文定义了Polarity Score和Volatility Score两个值。

首先，本文将交易日收盘价高于开盘价记作股价上涨，交易日收盘价低于开盘价记作股价下跌。

Polarity Score(PS)值用于评价一个词和股价上涨和下跌偏向性的关系。PS值越高，说明该词语越偏向于仅在股价上涨或者下跌日的新闻里出现。对于语料中的任意词 i 都有

$$PS_i = \frac{|n_{i,p} - rn_{i,n}|}{n_{i,p} + rn_{i,n} + k} \tag{12}$$

其中$n_{i,p}$为词 i 在对应股价上涨日的新闻中出现的次数，$n_{i,n}$为词 i 在对应股价下跌日的新闻中出现的次数，r 为对应股价上涨日新闻的总词数和股价下跌日新闻的总词数的比。k 为常数，目的是避免分

母为 0 的同时减少词频数过小的词的影响。理想状态下，仅在上涨日或者下跌日新闻中多次出现词的 PS 值会接近 1，而上涨日或者下跌日新闻出现次数差别并不明显词的 PS 值会接近 0。

Volatility Score（VS）用于值用于评价一个词和股价变动幅度的关系。VS 值越高，说明该词语越偏向于在股价激烈变动日的新闻里出现。对于语料中的任意词 i 都有

$$VS_i = \frac{\sum n_{i,j}(\Delta p_j - \Delta \mathrm{p} + 1)^2}{\sum n_{i,j}} \tag{13}$$

其中$n_{i,j}$表示词 i 在对应第 j 天的新闻中出现总次数，Δp_j表示在第 j 天股价的变动率的绝对值，$\Delta \mathrm{p}$ 则表示所有样本的股价变动率绝对值和的平均。理想状态下，更倾向于在股价变动率大于平均变动率的日期出现词的 VS 值会显著大于 1，而更倾向于在股价变动率小于平均变动率的日期出现词的 VS 值会接近于 0。

最终词 i 的特征值$PSVS_i = PS_i \times VS_i$。对于一个含有 d 个词的句子 s，其特征值$PSVS_s$ 为

$$PSVS_s = \frac{\sum PSVS_i}{d} \tag{14}$$

由定义可知，PSVS 值越高的词，越偏向于仅在股价激烈上涨或者价激下跌日的新闻里面出现，而一个句子的 PSVS 值越高，则表明句子中这类词的比例越高，则越有可能是对于股价预测有特征量的句子。

通过 PSVS 值，对于一整天的所有新闻，将语料拆分为一个一个句子然后分别计算他们的得分，按照得分子多少排序，依据模型所需要的文本最大长度，抽取在长度内排名更高的句子，截去排名靠后且超过文本最大长度的句子，以达到集中特征量的同时排除噪声的目的。

4 实验设置

4.1 实验目的

本文所设置实验的目的主要有以下两点：

1. 检验在分析超长文本新闻语料预测股市任务里，不同的分类模式之间是否有差距，最新的预训练模型是否一定优于传统的统计模型。
2. 检验 3. 2 中所提出的文本提取算法是否能够改善噪声多的问题。

4.2 预测对象及语料

本文所选定的预测对象为反映美国股市的标准普尔 500 指数（Standard & Poor's 500 INDEX）。

标准普尔 500 指数，是包含美国 500 家上市公司股价的一个股票指数。这个股票指数由标准普尔公司创建并维护，覆盖的所有公司都是在美国主要交易所，如纽约证券交易所、Nasdaq 交易的上市公司。与道琼斯指数相比，标准普尔 500 指数包含的公司更多，因此风险更为分散，能够反映更广泛的市场变化。

选择标普 500 指数作为预测的对象，是因为美股市场发展时间长，相对成熟，流通量大，信息量多，极端情况相对较少，而 sp500 作为美股市场的股价加权指数，能够进一步过滤各类极端情况，同

时因其涉及范围广，与金融相关的新闻或多或少都有所关联，语料搜集相对容易。而对于本文的实验目的来说，预测标普 500 指数已经足以满足验证的要求。

本文所搜集的标普 500 股价数据为从 2009 年 9 月至 2020 年 12 月期间的交易日的数据，来自于 investing. com 的公开数据，包括每一个交易日的开盘价，收盘价，最高值，最低值。总计包含 2, 776 个交易日，其中有 1, 430 个交易日收盘价高于开盘价记作上涨，1, 346 个交易日收盘价低于开盘价记作下跌，平均变动幅度 0. 76%。

本文所使用的新闻语料来自于路透社。路透社是是世界上最早创办的通讯社之一，是世界前三大的多媒体新闻通讯社，提供各类新闻和金融数据，在 128 个国家运行。作为一家有相当规模的媒体，其所报道的金融新闻的质量相对有保证的同时数量也不少。语料中包含 2009 年 9 月到 2020 年 12 月期间所有在路透社英文金融板块所公开的新闻，这些语料一部分来自于 Zhou 等[12]所公开的数据，一部分使用网络爬虫在路透社官网爬取，总计 189, 613 篇，平均每天约 68 篇，总计包含超过 300 万个句子。

4.3 数据处理

因为所有的文本数据来源于网页，在作为语料前还剔除 html 标签、特殊字符等处理，并根据不同的模式需要进行拆分或者合并等操作。

本文的预测目标是股价的变动，标签也由股价的变动决定。每一个日期的变动具体的定义为股价当日的开盘价减去收盘价，如果结果大于 0，则记作股价上涨，附予[1, 0]标签，如果结果小于等于 0，则记作股价下跌，附予[0, 1]标签。而每一个日期所对应语料的为前一交易日的开盘后至当前交易日的开盘前所发布的所有新闻。

文本提取的部分，本文使用每一个日期对应的全部新闻语料进行文本提取，并依据各模型各自的最大语料长度进行截取，截取后的语料同样基于当天的股价变动附予标签。

对于无法处理长文本的模式，本文也采取分篇预测的方式，最终的结果取决于各篇的预测结果上涨和下降哪一方多于另外一方。

4.4 训练集及测试集

训练集和测试集的划分中，通常的划分方式是分为多折并取其中一折作为测试集，而新闻语料具有时间关联，多折的方法会使同一折内多个同一时期的语料划作测试集，训练集却没有同一时期的数据，显然会影响模型的训练效果。

所以本文采取随机抽取测试集的方式，使测试尽量分布于多个时期。具体而言，每一轮的抽取，都会在总计 2, 776 个交易日的日期列表中以百分之 10 的概率抽取一个总数约为$10\% \times 2{,}776$个测试集日期列表，然后按照日期列表划分，测试集日期列表内的日期所对应的语料算作测试集，测试集日期列表内的日期所对应的语料算作训练集。在同一轮抽取中，不同的模型所使用的是同一个测试集日期列表。

4.5 实验环境及模型参数

本文所实验使用的平台为基于 Google Colab 提供的带有 GPU 的 Python 环境。

对于 BOW 和 TF-IDF 特征量的分类模式，本文使用基于 Sklearn 包的实现，词频最小过滤值 5，所对应的分类器使用 Sklearn 包默认参数实现，SVM 为 Linear Kernel。

对于 word2vec 特征量的分类模式，使用 word2vec 包构建，词向量训练语料为 wiki 官方公开的英文 wiki 语料，所对应的分类器使用 PyTorch 包构造，其中 rnn 隐藏层节点数 128，lstm 隐藏层节点数 128，所使用的优化器 Adam，初始学习率 10e-4，batch 大小为 16，最大句子长度 512。

对于 BERT-based 特征量的分类模式，使用 google 提供的 BERT base, hugging-face 提供的 RoBERTa-base 和 Araci 等提出的 FinBERT base 三种预训练模型，微调训练所使用的优化器 Adam，初始学习率 5e-5，batch 大小为 16，最大 token 长度 512。

4.6 评价标准

对于准确度并不高的股票预测任务，准确度是最优先的评价指标，而 F1 值在准确度并不高的股票预测任务中起伏较大，仅仅作为参考列出，不作为评价标准。例如在训练中有没有经过充分训练的分类器会将所有样本都分类为上涨或者下降，导致 F1 值达到 1.0 或者 0.0，但是实际上这样的分类器其实并没有分类能力。

其具体定义如下

表 4.1 预测结果和标签对照

预测\标签	上涨	下跌
上涨	TP	FP
下跌	FN	TN

则

$$准确率Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \tag{15}$$

$$F1\ 值 = \frac{2\times P\times R}{P+R} \tag{16}$$

其中$P = TP/(TP + FP)$，$R = TP/(TP + FN)$。

5 结果及分析

因为测试集和训练集是随机抽取，根据抽取结果的不同可能会有误差，所以本文的所有结果为抽取 3 次并分别训练测试的结果，最终值为平均值。

5.1 不同处理模式间的对比实验

表 5.1 不同预测模式的实验结果

模式	准确率	F1
Random guessing	0.50	0.50
BOW+LM	0.5186±0.0058	0.4563±0.2542
BOW+RF	0.5271±0.0073	0.5664±0.1323
BOW+SVM	0.5318±0.0067	0.5634±0.0462
TFIDF+LM	0.5323±0.0036	0.5024±0.0646
TFIDF+RF	**0.5413±0.0074**	0.5724±0.1757
TFIDF+SVM	0.5389±0.0053	0.5847±0.1424
Word2vec+RNN	0.5257±0.0034	0.5324±0.2662
Word2vec+LSTM	0.5314±0.0061	0.5526±0.1765
BERT-base	0.5377±0.0045	0.5664±0.1467
RoBERTa-base	0.5402±0.0053	0.6042±0.1289
FinBERT-base	0.5383±0.0032	0.5462±0.0921

从实验结果来看，没有使用任何外部信息特征的情况下，纯长文本语料的分类模式普遍准确率并不算高。其中，平均准确率最高的是 TFIDF+RF 这一模式。BOW 作为特征量的分类模式中，准确率最好的是使用 SVM 分类器，其次是 RF，最后才是分类能力相对较弱的 LM 模型。TF-IDF 作为特征量的分类模式中，RF 分类器最好，SVM 其次，最后是 LM 分类器。从表现形式的角度而言，TF-IDF 要好于 BOW。在使用 Word2vec 作为特征量的模式中，因为类 RNN 模型具有的长期依赖问题，并不具有很好的长文本处理能力所有准确率都不算高。在使用各类 BERT-based 模型作为特征量的模式中，标准的 BERT 模型和专门使用金融语料训练的 FinBERT 模型并没有太大差距，成绩最好的是 RoBERTa 模型。从表现形式的角度而言，TF-IDF 嵌入要好于 BOW 嵌入，而 BERT-based 嵌入要好于 word2vec 嵌入，BERT-based 嵌入和 TF-IDF 嵌入并没有显著差距。总体而言，现有的预训练语言模型相对于传统的统计模型，在处理新闻语料预测股价变动的任务上并没有显著的优势。

总体而言，现有的预训练语言模型相对于传统的统计模型，在处理新闻语料预测股价变动的任务上并没有显著的优势。

5.2 文本提取算法有效性实验

表 5.2 不同处理模式的实验结果

模式	准确率	F1
Random guessing	0.50	0.50
BERT-base	0.5377±0.0045	0.5664±0.1467
RoBERTa-base	0.5402±0.0053	0.6042±0.1289
FinBERT-base	0.5383±0.0032	0.5462±0.0921
PSVS+BERT-base	0.5675±0.0065	0.5735±0.0967
PSVS+RoBERTa-base	**0.5758±0.0038**	0.5931±0.1361
PSVS+FinBERT-base	0.5676±0.0047	0.5885±0.08517

从实验结果来看，在不使用使用任何外部信息特征的同时不改变模型输入输出形式的情况下，本文所提出的基于 PSVS 值的文本提取算法能够平均增加预训练模型约 3%的准确率。

表 5.3 BERT-base 模型每轮 epoch 训练平均使用时间

模式	时间(秒)
BERT-base	1243.61
PSVS+ BERT-base	22583.43

同时因为仅需要分析提取后的文本，模型的训练时间也大幅度缩短。

6 结论

本文通过搭建几种常见的文本处理模型进行实验对比，并提出了一种基于词频及股价变动量用于解决预训练语言模型的超长文本处理问题的文本提取算法。经过实验，现有的预训练语言模型相对于传统的统计模型，在处理新闻语料预测股价变动的任务上并没有显著的优势。而本文提出的文本提取算法则能够提高预训练语言模型约 3%的准确度的同时还能够大幅度缩短模型训练所需时间。对于股市预测任务来说，更加准确的特征量和标记或许比分析模型本身的性能更能影响结果。对于噪声多的长语料来说，处理这部分噪声不仅能减少计算量，还能够提高准确率。

参考文献：

[1] Jones, K.S.(1972). A statistical interpretation of term specificity and its application in retrieval. Journal of documentation.

[2] Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. CoRR, abs/1301.3781.

[3] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is All you Need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), Advances in Neural Information Processing Systems (Vol. 30).

[4] Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep Contextualized Word Representations. In M. A. Walker, H. Ji, & A. Stent (Eds.), Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, June 1-6, 2018, Volume 1 (Long Papers) (pp. 2227–2237).

[5] Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (cite arxiv:1810.04805Comment: 13 pages)

[6] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach (cite arxiv:1907.11692)

[7] Araci, D. (2019). FinBERT: Financial Sentiment Analysis with Pre-trained Language Models. ArXiv. /abs/1908.10063

[8] Schumaker, R.P. and Chen, H. (2009) Textual Analysis of Stock Market Prediction Using Breaking Financial News: The AZFin Text System. ACM Transactions on Information Systems (TOIS), 27, 12.

[9] Ding, X., Zhang, Y., Liu, T., & Duan, J. (2014). Using Structured Events to Predict Stock Price Movement: An Empirical Investigation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1415–1425). Association for Computational Linguistics.

[10] Pagolu, V., Reddy, K., Panda, G., & Majhi, B. (2016). Sentiment analysis of Twitter data for predicting stock market movements. In 2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES) (pp. 1345-1350).

[11] Hu, Z., Liu, W., Bian, J., Liu, X., & Liu, T. (2017). Listening to Chaotic Whispers: A Deep Learning Framework for News-oriented Stock Trend Prediction.

[12] Zhou, Y., & Voigt, K. (2020). Stock Index Prediction with Multi-task Learning and Word Polarity Over Time. ArXiv. /abs/2008.07605

[13] Chen, Q. (2021). Stock Movement Prediction with Financial News using Contextualized Embedding from BERT. ArXiv. /abs/2107.08721

[14] Lin, W., Tsai, C., & Chen, H. (2022). Factors affecting text mining based stock prediction: Text feature representations, machine learning models, and news platforms. Applied Soft Computing, 130, 109673. https://doi.org/10.1016/j.asoc.2022.109673

[15] Li Y, Pan Y. (2022). A novel ensemble deep learning model for stock prediction based on stock prices and news. Int J Data Sci Anal.13(2):139-149. doi: 10.1007/s41060-021-00279-9.

[16] Villamil, L., Bausback, R., Salman, S., Liu, T. L., Horn, C., & Liu, X. (2023). Improved Stock Price Movement Classification Using News Articles Based on Embeddings and Label Smoothing. ArXiv. /abs/2301.10458

[17] Ashtiani, M. N., & Raahemi, B. (2023). News-based intelligent prediction of financial markets using text mining and machine learning: A systematic literature review. Expert Systems With Applications, 217, 119509.

[18] Rong, X. (2014). Word2vec Parameter Learning Explained. ArXiv. /abs/1411.2738