自然语言处理与经济学研究范式变革：一个文本分析、理解与生成的框架

杨国超　院　茜　龚　强*

摘　要：自然语言处理（NLP）是一种依靠计算机和算法对人类语言进行分析、处理和计算，从而实现自动理解甚至再次生成自然语言的全新方法。本文首先总结了相关技术方法的演变过程，然后基于NLP方法的发展历程，依次介绍包括 Word2Vec、BERT 以及大语言模型的原理，并从分析文本、理解文本和生成文本三个方面，总结 NLP 方法在学术研究中的应用框架。NLP 技术正推动经济学研究范式实现三重跃迁：从依赖结构化数据到非结构化文本，从表层特征量化到深层语义理解，从解释过去转向模拟未来。

关键词：自然语言处理；机器学习；大语言模型
DOI：10.13821/j.cnki.ceq.2026.03.15

一、引　　言

自然语言处理（Natural Language Processing，NLP）技术的兴起与成熟，正成为经济学研究应对数据基础发生根本性变革的关键驱动力。在数字经济时代，经济学研究所依赖的数据，其形态、粒度与生成方式正发生深刻变化：第一，从以结构化数值数据作为分析要素的时代向包括财经新闻、政策文本、社交媒体等海量的非结构化文本数据的时代全面转变。第二，从以年月为单位的低频数据向以时分秒为单位的高频甚至即时文本数据转变。第三，数据生成方式也从仅由人类生成拓展至由机器自动生成。这些变化使得传统局限于结构化数据处理的计量方法面临瓶颈，而能够对文本进行自动分析、理解与生成的自然语言处理技术，则为经济学家系统利用这些新型数据、拓展研究边界、回答新的研究问题提供了不可或缺的工具，进而推动经济学研究范式发生深刻变革。为

* 杨国超，复旦大学保险应用创新研究院、复旦大学创新与数字经济研究院；院茜，河南大学高级金融学院；龚强，中南财经政法大学文澜学院。通信作者及地址：龚强，湖北省武汉市中南财经政法大学南湖校区，430073；电话：027-88386678；E-mail：qianggong@zuel.edu.cn。本研究得到国家社科基金重点项目（22AGL013）、国家社科基金重大项目（24ZDA032）、国家社科基金重点项目（23AZD029）及国家自科基金面上项目（72073146）的资助。作者感谢北京大学刘诗尧、匿名审稿人和期刊主编的宝贵建议，当然文责自负。

此，本文通过梳理自然语言处理方法在现有学术研究中的应用逻辑，以探讨学术界如何借助这些方法提出、分析并解决问题，同时为未来研究提供有益的启示与指导。

现有综述类文献集中于讨论文本分析在研究中的应用（Li，2011；Loughran and Mcdonald，2016；唐国豪等，2016；沈艳等，2019；姚加权等，2020；Bochkay et al.，2023），而随着自然语言处理技术的推陈出新，许多文本分析技术已经发生显著改进，例如考虑主题情感或观点的主题分类模型、考虑向量长度的文本相似度模型等；随着大语言模型的出现，文本分析逐渐被更高阶的自动化语言理解和生成所取代。本文在系统性回顾自然语言处理技术发展历程的基础上，进一步介绍了包括 BERT、GPT 和以 DeepSeek 为代表的新一代大语言模型，并探讨其在研究中的应用逻辑，从而为推动研究方法的变革提供一个前瞻性的框架。

此外，现有文献还侧重于从技术视角对新的技术方法展开综述（Gentzkow et al.，2019；马长峰等，2020；李春涛等，2024），而同时更为重要且更为基本的问题是，新技术的应用逻辑也值得被关注。本文在介绍代表性的技术方法外，也着眼于自然语言处理在学术研究中的应用逻辑，从分析文本、理解文本和生成文本三个方面总结出一个切实有用的基于自然语言处理方法的应用框架，以适应未来经济学研究范式的变革。具体地，文本分析作为量化文本特征的工具，已逐渐一般化为研究中的常用方法；文本理解未来会更多地被用于更精准地识别作用机制，成为打开理论“黑箱”的有力工具；生成式大语言模型的不断发展也为文本生成的潜在应用场景带来了更多机会。

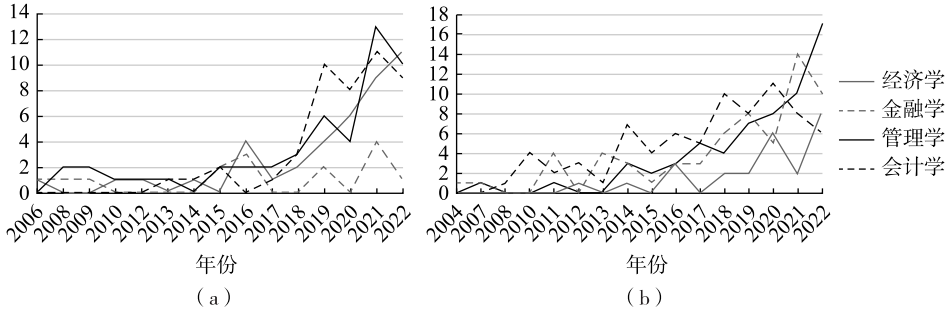
二、现有文献统计及研究方法变迁

本文以“文本分析”“自然语言处理”“NLP”等 29 个关键词在中国知网、Web of Science 以及 Elsevier 数据库进行文献检索，并辅以人工筛选和阅读，共搜集到发表在权威期刊上的中文论文 139 篇，英文论文 205 篇。[①] 图 1 显示，英文文献比中文文献更早地开始利用自然语言处理方法，从总体趋势来看，2017 年之后中英文发表均显著增加。

从研究方法来看，由于词典法简单易行，因此被最早也最广泛地应用到学术研究中，此后，主题模型等机器学习方法的应用也越来越多。总的来看，英文文献比中文文献更早地开始使用自然语言处理方法开展研究，且各类研究方法在英文文献中的应用平均比中文文献早约 5 年。图 2 绘制了中英文文献中所

① 检索过程及结果统计详见附录Ⅰ。由于篇幅所限，附录未在正文列示，感兴趣的读者可在《经济学》（季刊）官网（https://ceq.ccer.pku.edu.cn）下载，或联系作者获取未删减版全文。

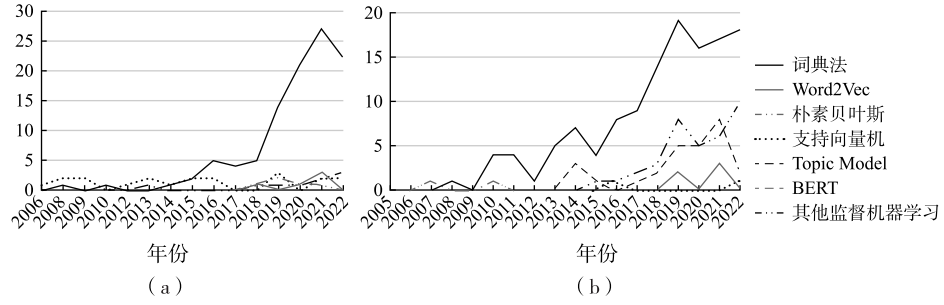
使用自然语言处理方法的总体变迁图。



（a）　（b）

图 1　中英文期刊中相关文献的发表趋势

注：图 1(a)为自然语言处理方法在各学科中文期刊中的发表数量变化趋势，图 1(b)为对应英文期刊中的变化情况。



（a）　（b）

图 2　中英文期刊中相关文献的方法变迁

注：图 2(a)为各类自然语言处理方法在中文期刊中的发表数量变化，图 2(b)为对应英文期刊中的变化情况。如果一篇文献同时使用了多种自然语言处理方法，则统计时会对该文献进行重复计数。

学术研究中采用的自然语言处理方法始终随着 NLP 技术的进步而持续演进。从早期的基于统计的字典法到基于机器学习的主题分类模型，再到深度学习方法推动下的词嵌入技术，乃至如今超大规模参数的大语言模型，传统方法也在技术迭代中不断发展。表 1 列示了在不同应用场景下自然语言处理方法的拓展演变过程。

表 1　传统自然语言处理方法的发展和演变

NLP 技术		传统方法	拓展演变
文本表示方法		词袋法	→考虑单词重要程度的 TF-IDF→考虑上下文的词嵌入方法→具有复杂语义表征能力的大语言模型
文本相似度	向量表示	词袋法	→Word2Vec 等词嵌入法→BERT 等基于大模型的文本表示方法
	相似度计算	余弦相似度等仅考虑向量夹角的计算方法	→向量相似度等同时考虑向量长度和向量夹角的计算方法

（续表）

NLP 技术	传统方法	拓展演变
命名实体识别	基于既定规则的识别方法	→基于机器学习的识别方法→基于大模型的识别方法
文本主题分类	以 LDA 为代表的概率模型	→以 BERTopic 为代表的语义聚类方法
词嵌入方法	以 Word2Vec 为代表的局部上下文感知模型	→以 GloVe 为代表的全局建模方法
大语言模型	以早期 ChatGPT 为代表的初代大语言模型	→以 DeepSeek 为代表的推理式新一代大语言模型

三、自然语言处理的发展阶段及方法梳理

自然语言处理的发展分为四个阶段：第一阶段的 NLP 方法主要基于统计思想，实现对自然语言的初步统计与分析。第二阶段的 NLP 方法将任务转化为机器学习中的分类任务，但在流程设计和特征构建中仍需人工干预，导致模型在泛化能力和运行效率方面存在缺陷。第三阶段的深度学习方法则借助深度神经网络，利用大规模语料库和丰富的上下文信息，实现对自然语言更深入的理解。这类方法在预测准确率、预测效率和泛化能力等方面均取得了显著提升。随着 Transformer 框架的提出（Vaswani et al.，2017），NLP 进入第四阶段大语言模型时代。大语言模型基于大规模数据训练得到拥有大量参数的通用预训练模型，在实践中仅需零样本或少样本学习即可灵活而高效地处理复杂任务。大语言模型在多个 NLP 任务中取得出色表现，也正在重塑经济学的研究范式。

（一）基于统计的自然语言处理方法

1. 以词典法为起点的文本表示方法及其演进

词典法是经济学和管理学研究中最常用且最基础的一种文本分析方法，该方法基于文本向量的表示方法——词袋法发展而来。词袋法将一段文本转化为向量表示，单词出现即表示为 1，否则为 0。然而，该方法无法捕捉到单词的含义，导致无法识别同义词、同词不同义等重要的语义现象。此后，基于深度学习的词嵌入方法相继出现，如 Word2Vec、GloVe、FastText 等，这些方法通过利用单词的上下文信息生成包含语义特征的词向量表示，能够更好地捕捉到单词含义及单词之间的语义关系（Li et al.，2021）。

恰当的文本预处理是许多 NLP 应用中的关键环节，直接影响模型的输出质量，详细的文本预处理步骤见附录Ⅱ。但近年来，基于 Transformer 框架发展而来的许多大语言模型（如 BERT、GPT、DeepSeek 等）无需对原始文本进行

复杂的预处理或向量化表示，就可以直接处理句子和文本。Transformer 框架中的自注意力机制使模型能够在上下文中动态捕捉词语之间的依赖关系，实现上下文敏感的文本理解（Vaswani et al.，2017）。这一特点使得大语言模型在处理诸如社交媒体数据、评论数据等表达灵活、语义复杂的文本时具有独特的优势，弥补了词典法、Word2Vec 及 GloVe 等静态词向量表示方法的不足。

2. 文本相似度：从词袋模型到语义匹配，从向量夹角到向量长度

文本相似度是利用统计方法计算两个文本间相似程度的自然语言处理方法，研究中常利用文本相似度刻画信息披露的“样板化程度”。文本相似度的计算分为以下两个步骤：第一，向量表示，通常利用特征提取方法将经过预处理的文本表示为数值型向量，例如词袋法、Word2Vec 等词嵌入方法以及 Sentence-BERT 等基于大模型的文本表示方法。第二，计算相似度，选择恰当的统计方法计算向量间的相似程度，例如余弦相似度、杰卡德相似度、最小编辑距离相似度等。相比于传统方法，Sentence-BERT、USE（Universal Sentence Encoder）等基于大模型的文本表示方法会通过将整个句子或短语表示为一个固定长度的向量，有效表达句子的整体语义信息，从而在解决句子级别的语义相似度比较、语义检索等方面表现更好。

尽管余弦相似度是研究中最常用的相似度算法，但其存在一个明显缺陷，即只要两个向量的夹角相同，即使向量长度差异很大，计算得到的余弦相似度仍然相同。这是因为该方法仅关注向量的方向，而并未考虑文本的长度。Srivastava（2023）提出的向量相似度（vector similarity metric）则同时考虑了向量的夹角和长度差异，从而能够更全面地衡量两个向量间的相似度。[①] 如图 3 所示，在分别计算两组向量（$\boldsymbol{A}$，$\boldsymbol{B}$）和（$\boldsymbol{A}'$，$\boldsymbol{B}$）的文本相似度时，$\boldsymbol{A}$ 与 $\boldsymbol{A}'$ 的方向相同但长度不同，因此使用余弦相似度会得到相同的结果；而向量相似度在考虑长度差异后会得到不同的相似度值，从而能够更准确地区分两组向量之间的差异。

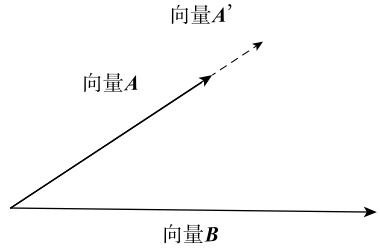


图 3　向量相似度的度量方法示意图

① 余弦相似度的计算公式为 $diff_{A,B}=1-\frac{Vector_A \cdot Vector_B}{|Vector_A|\cdot|Vector_B|}$，向量相似度的计算公式为 $diff_{A,B}=1-\frac{|Vector_A-Vector_B|}{\sqrt{|Vector_A|^2+|Vector_B|^2}}$。

3. 命名实体识别：从既定规则到机器学习，再到大语言模型

命名实体识别是 NLP 领域中最基础的任务之一，用于从文本中识别和定位具有特定意义的实体，例如人名、地名、机构名、日期、货币、时间等。现有研究将文本中的实体信息占比称为文本的“具体性”程度或“硬信息”占比。李晓溪等（2019）发现被监管问询后，企业显著提升了其并购重组报告书中的文本“具体性”程度，表明监管问询通过改善信息披露缓解了并购交易中存在的信息不对称问题，从而发挥了监管作用。

早期的命名实体识别方法主要依赖于预定义规则和词典来识别已知的命名实体，这就导致其缺乏对新实体的发现能力。不同于仅能基于给定规则识别命名实体的词典法，随后出现的统计机器学习模型，如条件随机场、隐马尔可夫模型等，通过序列建模和特征工程，有效捕捉单词之间的依赖关系，从而判断一个单词是否属于某个实体类别，极大地提高了命名实体识别的灵活性和准确性。深度学习模型如双向长短期记忆网络则通过双向学习输入序列前后的信息，提升了未知命名实体的发现能力（Huang et al.，2015）。近年来，基于 Transformer 架构的理解型大语言模型，如 BERT、RoBERTa 等，不仅能够更有效地捕捉复杂的语义关系，还通过自监督学习和迁移学习等技术，增强了其在无标注数据环境下的自适应实体识别能力。

（二）基于机器学习的自然语言处理方法

早期研究多使用朴素贝叶斯、支持向量机等有监督的机器学习对文本进行分类，具体介绍见附录Ⅲ。尽管这类方法在分类标签和类别数量已知的情况下表现良好，但在面对内容复杂、类别数量未知或难以预设的文本时却无能为力。基于无监督机器学习发展而来的 LDA、BERTopic 等方法则无需人工确定分类标签和分类数量，而是通过对文本特征的自动识别与聚类，实现对文本的主题分类。

LDA 模型本质是一种概率生成模型，该模型假定主题与单词之间的分布满足共轭的狄利克雷分布，即假设文档中出现的单词是由一组隐含的主题所生成，并且主题与单词的分布参数满足狄利克雷分布（Blei et al.，2003）。LDA 模型假定包含 m 个文本的语料库 M 中存在 D 个主题，首先通过计算困惑度指数（perplexity score）确定最佳主题个数，然后对经过预处理的语料库 M 进行无监督学习，最终 LDA 模型会以概率分布的形式给出每个文本的主题分布，并且每个文本所包含的单词将按照这些单词与不同主题的相关性被重新分组，通常按照相关性强弱列示 10 至 15 个单词即可大致了解各主题的含义。换句话说，利用 LDA 主题模型能够获取多个文本的主题、各主题的关键词分布，以及每个文本中各主题的概率分布，从而实现对文本主题的自动化识别与分类。

目前主题模型已发展出许多变体，包括相关主题模型（Correlated Topic

Model，CTM）、动态主题模型（Dynamic Topic Model，DTM）、监督主题模型（Supervised Latent Dirichlet Allocation，sLDA）等。[①] 例如，Donovan et al.（2021）利用 Blei and Mcauliffe（2007）提出的监督主题模型考察了公司电话会议文本中与债券风险相关的主题，sLDA 模型会根据人工对文档的标注，在对文档进行主题分类时生成与标注相关的主题，从而使得所得的主题分类更具可解释性。Chen and Mankad（2024）还开发了结构主题与情感论述模型，其能够在文本主题分类的基础上捕捉每个主题的情感或观点，修正了传统主题分类方法无法进行统计推断的缺陷，特别适用于情感色彩明显的文本分类任务，如博客、评论数据等。

不同于 LDA 模型依托于“词频统计＋概率模型”的主题分类思想，随后出现的 BERTopic 模型则采取语义聚类的思路，充分发挥 BERT 预训练语言模型在语义理解方面的优势，并结合降维与无监督聚类算法，从而能够应对更复杂语义场景下的主题识别与分类任务。具体地，BERTopic 首先利用 BERT 模型将文本转化为具有上下文感知能力的高维语义向量；其次，通过 UMAP 等降维算法将高维语义向量映射到可聚类的低维空间，降维算法能够在保留语义结构的前提下大幅压缩向量维度，从而缓解维度灾难并为聚类做准备；随后，使用 HDBSCAN 等基于密度的无监督聚类方法，将语义相近的文本自动归类为若干簇，无需预先设定主题数量；最后，采用 c-TF-IDF 的权重计算方法从每个簇中抽取最具区分度的关键词，这些关键词有助于快速理解文本所涵盖的主题及其含义。表 2 列示了以 LDA 为代表的传统主题模型与以 BERTopic 为代表的基于深度学习的主题模型的对比。

表 2 LDA 模型与 BERTopic 模型的区别

	LDA 模型	BERTopic 模型
基本原理	词频统计＋概率分布	深度语义理解＋密度聚类
主题数量	无需预先设定	可以预先设定也可以不设定
优势	计算成本低；适用于表达较为规范的长文本	具有较强的上下文语义理解和表示能力，能够处理一词多义问题
劣势	难以理解上下文；对短文本等稀疏语料的分类效果较差	所需计算成本较高
适用场景	政策文本、学术论文等表达规范的长文本	社交媒体文本、网络评论文本、多语言文本等表达灵活多变的短文本

① 相关主题模型适用于文本所包含主题之间存在相关性的情形，克服了 LDA 模型假设主题相互独立的局限性；动态主题模型在原有 LDA 框架的基础上增加了时间序列的考虑，能够对文本主题进行动态演变分析，适用于内容变化迅速的社交媒体文本、长周期文本等；监督主题模型则将监督学习引入主题生成过程，从而能够用于特定任务的主题识别。韩亚楠等（2021）对基于 LDA 模型的各种拓展模型进行了详细介绍。

（三）基于深度学习的自然语言处理方法

1. 神经网络

深度学习是机器学习发展过程中最重要的分支，而神经网络则是深度学习的核心和基础。神经网络通过模拟大脑神经元的工作机制，利用多层结构逐步提取数据特征，最终实现预测、分类等任务。具体地，每个神经元接收到输入数据后，通过加权求和和非线性激活函数对数据进行处理，输出结果再传递给下一层神经元。多个层次的神经元共同构成了深度神经网络结构（如图 4 所示）。这一过程类似于人脑处理信号的方式，通过不断调整权重参数，神经网络可以自动从数据中学习复杂的特征和模式，实现从简单输入到复杂输出的映射。初代神经网络即感知器模型仅能够进行简单的线性分类，随后的 BP 算法通过引入多层感知器结构能够解决非线性问题，但在提取复杂特征方面仍然存在局限。卷积神经网络（Convolutional Neural Network，CNN）通过局部连接和权重共享机制，能够有效提取局部特征，显著提升了神经网络的表现能力，这种局部特征提取能力能够有效识别文本中的短语和关键词（Kim，2014；Obaid and Pukthuanthong，2021）。循环神经网络（Recurrent Neural Network，RNN）则通过捕捉序列数据的时间依赖性，能够更好地处理序列数据，例如利用其捕捉上下文信息，从而理解文本语义。

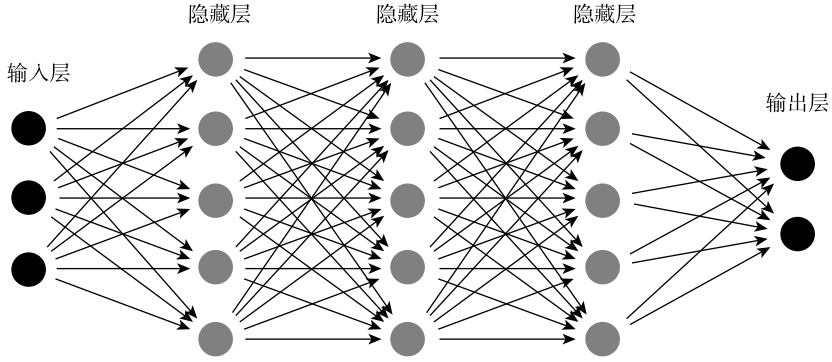


图 4 神经网络示意图

2. 词嵌入模型：从 Word2Vec 到 GloVe，从局部上下文感知到全局建模[①]

Word2Vec 是一种基于浅层神经网络的词嵌入方法，该方法将文本中的每个单词都转换为向量表示，即高维空间中的一个点，这个空间的维度总数与文本的总词数相同。词嵌入的过程则是通过将单词映射到实数向量，从而将高维空间转换为低维空间的过程，进而实现对单词语义信息的编码和表达。

① 针对 Word2Vec 和 GloVe 方法的实例介绍详见附录Ⅳ。

Word2Vec 通过将上下文词汇表示为连续的、分布式的向量来预测目标词汇。与词袋法等基于词频的文本向量表示方法相比，利用 Word2Vec 得到的词向量能够更好地反映单词在上下文中的语义关系。

实践中常常利用 Word2Vec 在特定语境下为指定单词寻找相似词，从而扩充关键词表用以构建文本变量。例如，姜富伟等（2021）基于 LM 词典（Loughran and Mcdonald，2011）选取了若干种子词汇，运用 Word2Vec 将多个来源的文本数据转化为空间向量表示，然后计算向量的余弦相似度找到与种子词高度相似的新词汇，从而构建一个适用于中文财经文本的情感词典。王靖一和黄益平（2018）基于海量的新闻媒体报道数据，结合 LDA 主题分类模型和 Word2Vec 方法构造了中国金融科技情绪指数。

不同于 Word2Vec 依赖于局部上下文推算单词的语义，GloVe 利用全局词共现统计信息反映单词之间的语义关系，其核心思想是，词与词之间共现概率的比值往往比共现概率本身更能揭示词之间的语义关系。具体地，GloVe 模型首先根据语料库构建单词的共现矩阵，用以计算每个单词在全局上下文中出现的频率；然后构建目标函数用以计算每个词向量，使其内积能够近似表达单词间的共现概率，以确保语义上越相近的词向量在空间上的距离越小，从而捕捉到更准确的语义关系。换言之，与 Word2Vec 局限于使用局部上下文窗口内的共现信息表示单词向量不同，GloVe 基于整个语料库的全局共现信息来构建单词向量，这种全局视角不仅能够捕捉单词与邻近上下文中的关系，还能更好地反映单词在整个语料中的相对重要性和语义位置，从而被用来理解更复杂的语义关系。

（四）大语言模型

大语言模型（Large Language Model，LLM）所采用的“预训练加微调”模式使其能够在无标注或少量标注数据的微调下达到较好的任务处理效果。具体地，预训练阶段通过对海量数据的自监督学习，使模型初步具备通用或特定领域的知识，微调阶段则通过少量标注数据对模型进行微调，使其快速适应具体任务需求，这不仅显著提升了模型的泛化能力和迁移学习能力，还降低了数据标注成本。大语言模型区别于过去深度学习模型的关键在于 Transformer 框架，该框架通过引入自注意力机制，克服了 RNN、LSTM 等传统深度学习模型在处理长序列数据和长距离依赖方面的局限性，且能够并行处理整个输入序列，并动态捕捉句子中任意两个词之间的依赖关系，不论这两个词的距离有多远（Vaswani et al.，2017）。换言之，与传统深度学习模型“逐字逐句”处理数据的方式不同，注意力机制通过模拟人类注意力分配的方式，能够“一目十行”快速捕捉到关键信息，极大地提升了模型的理解能力和计算效率。

基于 Transformer 框架的大语言模型不断衍生发展，目前可以被分为以下几类：第一类是以 BERT 为代表的理解型大语言模型，主要采用 Transformer 框架中的编码器结构（Encoder-only），在自然语言理解任务中表现出色；第二类是基于解码器（Decoder-only）的生成式大语言模型，具有代表性的模型包括 GPT、Kimi 等模型，这类模型能够兼顾文本理解和生成，擅长文本生成任务；第三类是基于编码器和解码器（Encoder-Decoder）的大语言模型，例如 T5、BART、GLM 等，这类模型采用双向注意力机制，能够同时进行文本理解与生成，但随着大语言模型的训练和发展，这类模型在泛化能力以及处理效率上提升有限，逐渐被 Decoder-only 模型所超越。以 DeepSeek 为代表的新一代大语言模型则进一步改进 Transformer 架构，通过引入诸如多头潜在注意力机制、稀疏注意力机制等革新的注意力机制，在计算资源有限的情况下能够更高效地理解和生成文本。大语言模型精准的理解能力和强大的表达能力再一次推动了经管研究的变革（钱贵明等，2025）[①]。

1. 以 BERT 为代表的理解型大语言模型

与传统的单向序列神经网络模型相比，BERT 通过对上下文的双向捕捉和全局依赖建模能够更深入地理解复杂文本，捕捉更丰富的语义信息。BERT 采用双向 Transformer 的 Encoder 模块，也被称为自编码（auto-encoding）模型。BERT 模型可以分为预训练（pre-training）和微调（fine-tuning）两个阶段。在预训练阶段，BERT 模型基于大量文本语料进行训练[②]，采用无监督机器学习方法完成两个子任务，分别是遮蔽语言模型和下一句预测模型。经过预训练的 BERT 模型已经编码了充足的语义信息，在实际应用中，可直接通过引入小规模的标注语料对 BERT 模型进行微调，以完成特定的自然语言处理任务。例如，Lee and Zhong（2022）首先基于中国投资者在交流平台与上市公司的对话文本构建语料库，由人工根据文本的性质和内容对文本进行分类标注，然后使用该标注样本对 BERT 模型进行微调，最终将投资者发帖数据根据文本性质和内容进行分类。

尽管 BERT 在通用领域表现出色，但实践中若想要在特定专业领域取得优异表现，就需要研究人员使用专门数据进行微调。为了得到更适用于金融领域的大语言理解模型，研究者开发了针对金融专业领域的 BERT 模型。例如，Huang et al.（2022）利用海量财经专业文本进行预训练得到 FinBERT 模型，并使用情感分类任务检验模型的有效性，结果表明，FinBERT 模型的表现明显优于词典法、朴素贝叶斯法以及基于 Word2Vec 的词嵌入法等在情感分类中的表现。Zhang et al.（2021）针对财经类新闻和研究报告数据，训练了 Mengzi 模型

① 本文还对大语言模型的评估方法进行了梳理，详见附录Ⅴ。

② 包括 Wikipedia 和 BooksCorpus 等数据。

用于处理中文文本。与通用语言模型相比，该模型具有轻量和高效的特点，在处理专业领域的文本时表现更好。姜富伟等(2024)基于 BERT 模型训练出应用于中国金融市场情景的大语言模型，发现相较于词典法，该模型在预测金融市场回报等方面表现优异。

2. 初代生成式大语言模型[①]

与 BERT 不同，GPT 模型采用单向 Transformer 的 Decoder 模块，即去掉了第二个多头注意力机制(Multi-Head Attention)，仅保留了带掩码的多头注意力机制(Mask Multi-Head Attention)，也被称为自回归(Auto Regressive)模型。GPT 模型的训练过程可以分为基础大模型训练、指令微调和基于人类反馈的强化学习三个阶段。第一阶段模型基于大规模语料通过自监督学习完成预训练。第二阶段通过将多个不同任务整合到一个统一的生成式框架中构造“指令-答案”的训练语料对模型进行微调。这一阶段训练得到的模型能够根据指令生成合理的答案，但由于缺乏对人类偏好和语言习惯的深度理解，其表述仍与人类的语言存在差异；第三阶段，即在反馈学习阶段，采用基于人类反馈的强化学习方法，通过与人类的持续交互学习，最终生成符合人类常识、逻辑和语法的自然语言。

然而，生成式模型仍然面临许多挑战。其一，由于在训练过程中错误地使用未来数据而导致前视偏差问题，该问题在实践中可以使用匿名数据或按照时间训练数据(Drinkall et al., 2024)加以缓解。其二，模型在生成内容时会出现“幻觉”问题，即输出看似合理但与事实不符或存在错误的内容，实践中可通过策略性微调或事实校验工具加以解决。比如，Krishna et al.(2024)开发的 GenAudit 工具缓解了该现象。其三，模型输出的不确定性，实践中可通过多次运行与参数控制等提高输出的稳定性和可重复性。其四，在使用模型模拟人类行为的实验中，模型可能因曾经阅读过相关文献而出现确认偏差现象。可以通过设计对照实验、采用多个不同模型对同一任务进行独立分析等方法加以缓解。最后，值得一提的是，检索增强生成通过从外部知识库检索相关信息并整合至用户输入，有效提升了生成结果的事实准确性和时效性，同时降低了微调成本和潜在的数据泄露风险，从而常被应用于构建专业或私有的智能问答系统。

3. 新一代推理式大语言模型

尽管以 ChatGPT 为代表的初代大语言模型引发了学术界和工业界的广泛关注，但其运行中出现的幻觉现象、复杂推理能力不足以及对算力的过度依赖

① 指 GPT-4 之前的模型。

等问题，使得模型发展陷入瓶颈。事实上，这类模型的发展策略可以归纳为两点，规模驱动和数据驱动。前者通过不断扩大规模、无限增加参数量的方式提升模型性能，即采取“大力出奇迹”的发展模式，导致模型训练成本高企；后者依赖海量语料训练以优化模型的生成质量，即认为“读书破万卷，输出如有神”，但随着高质量可训练数据的逐渐耗尽，模型的性能提升也面临瓶颈。相较之下，以 DeepSeek、GPT-4、Claude3 等为代表的新一代大语言模型，通过引入链式思维（Chain of Thought，CoT），不仅显著增强了模型的复杂推理与逻辑思考能力，更从根本上推动了大语言模型从“知识生成”向“推理求解”的核心范式演进。

具体地，首先，与初代大语言模型依赖人类反馈强化学习不同，推理式大语言模型充分发挥强化学习的优势，通过自我迭代优化机制激发模型的长链思考和复杂推理能力，使大语言模型逐渐从过去的“快反应”问答模式转变为“慢思考”推理模式，不仅显著提升了模型在逻辑推理、数学推导、编程等复杂问题中的解决能力，减少了幻觉现象和错误生成，而且有望突破大语言模型对大规模训练数据的依赖瓶颈。同时，模型自我思考能力的提升降低了对提示语的依赖，有效改善了过去大语言模型输出过于依赖提示语的困境。当前，新一代 DeepSeek 模型还创新性地融合了快速响应的“非思考模式”与深度推理的“慢思考模式”，提出混合推理架构，在使用中可以根据需要自由切换模式，从而提升了使用的灵活性以及对算力的利用效率。

其次，推理式大语言模型采取的混合专家机制（Mixture of Experts，MoE）显著提升了模型的运行效率。这种模式借鉴类似于“分而治之”的策略，将模型划分为擅长处理通用任务的共享专家和多个专注于特定领域的专业专家，通过动态路由机制根据不同处理任务特点，将任务分配给最相关的专家，不仅能够有效解决任务分配不均的问题，而且通过稀疏激活模式显著降低了计算能耗。

最后，推理式大语言模型通过信息压缩、稀疏计算等方式优化了长文本理解能力，同时大幅降低了计算成本。以 DeepSeek 采用的多头潜在注意力（Multi-Head Latent Attention，MLA）机制为例，该机制由两部分组成：其一，潜变量模式改变了传统方法直接存储输入序列全部信息的方式，采用“信息压缩”策略将原始输入序列隐式编码为更高层次的语义特征，这种方式不仅增强了模型对长文本上下文语义的捕捉能力，还显著减少了存储和计算开销；其二，多头注意力机制通过并行计算多个注意力头，使模型能够根据不同子任务需求动态调用相同的语义特征，从而以并行且灵活的方式处理复杂任务。

四、自然语言处理在经管学术研究中的应用逻辑

本部分从分析文本、理解文本和生成文本三个方面，探讨自然语言处理在学术研究中的应用逻辑，并基于此构建出一个切实有用的基于自然语言处理方法的分析框架，如图 5 所示。

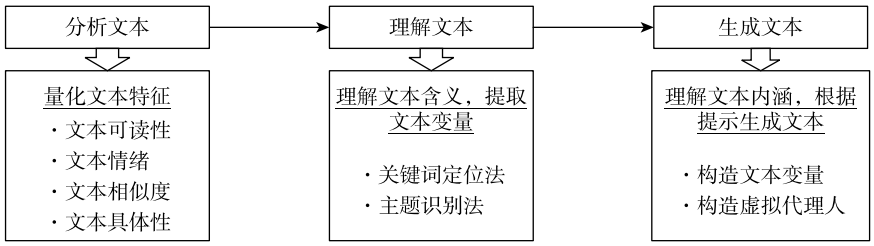


图 5　自然语言处理在经管学术研究中的应用逻辑

（一）分析文本

分析文本是 NLP 方法在研究中应用的初级阶段，它通过量化文本的关键特征，如文本可读性、文本情绪、文本相似度和文本具体性等，将传统的以结构化数据为主要研究对象的研究范式全面拓展至非结构化文本数据，解决了过去由于关键变量难以刻画而无法被回答的研究问题。例如，Dyer et al.(2017)使用 LDA 主题模型对 1996—2013 年间美国上市公司披露的年报进行主题分类，并从文本可读性、文本长度、文本样板化程度等多个角度对不同主题的信息披露内容展开分析，全面刻画了美国上市公司年报信息披露的内容及特征变化。Cohen et al.(2020)通过探究企业年报文本的纵向文本相似度与公司股价变动之间的关系发现，公司年报的价值相关性并未随着年报披露长度的增加而减弱。鉴于 Li(2011)、Loughran and Mcdonald(2016)、唐国豪等(2016)、沈艳等(2019)、马长峰等(2020)等已对多种文本分析方法及其在研究中的应用进行了详细的梳理与总结，下文将重点介绍理解文本和生成文本这两种应用逻辑。

（二）理解文本

文本分析帮助我们从某些特定角度对文本特征形成初步认识，但仍无法全面理解文本，即文本究竟在讲什么。文本理解则利用更为复杂的 NLP 方法，如 Word2Vec、主题模型、BERT 等方法，深入理解文本的具体内涵并提取相关变量，从而解决过去利用结构化数据无法观测到，或因存在测量偏差而无法有效解决的研究问题。

1. 关键词定位法——以管理层短视行为为例

关键词定位法首先通过专家或辅以词嵌入技术确定若干描述研究对象的关键词，然后通过计算这些关键词在文本中的词频占比等属性构建文本变量。以管理层短视行为为例，早期研究采用短期投资占比等财务指标刻画管理层短视行为，但这类变量仅能从事后捕捉到管理层短视行为的特征，而无法直接刻画管理层短视行为。事实上，管理层短视是管理层的一种认知行为，而语言作为人类认知的直接载体，更能反映人的认知、偏好和个性，通过分析语言中使用的词语类型和词频来捕捉人的特质，为直接刻画管理层短视行为提供一个可行的思路。为此，Brochet et al.(2015)基于人工阅读企业电话会议文本，首先确定了 7 个反映企业管理层短视主义的关键词[①]，然后通过计算这些关键词在电话会议文本中的占比构建了管理层短视主义指标。

2. 主题识别法——以企业创新行为为例

主题识别法通常利用主题模型识别与研究对象相关的主题内容，并通过计算这些文本与总文本的占比等属性来构建文本变量。早期研究利用研发投入或专利数量衡量企业的创新行为，但这些指标仅能够反映企业的技术或产品创新，却难以捕捉其他形式的创新活动，如新的生产方法、新市场开发、新组织形式等。为此，Bellstam et al.(2020)利用 LDA 对 1990—2021 年间标普 500 的分析师报告文本进行主题分类，提取创新相关主题并计算其文字占比，发现该文本指标能够全面捕捉到不同类型的创新活动：对于专利企业而言，其能够识别出具有较高价值的创新活动；而对于无专利企业而言，其能够识别出其他形式的企业创新活动。

为节约篇幅，附录Ⅵ还梳理了使用 NLP 方法进行理解文本的更多应用场景。

（三）生成文本

生成式人工智能工具能够基于对文本内涵的理解在不同情境下生成逻辑连贯、符合语言规范的文本输出，这种强交互性的特点，使得过去仅能由人工生成的内容如今可以借助自动化方法实现。

1. 利用文本生成方法构造文本变量

与文本理解类方法需要训练或微调，并对模型输出进行一系列转化计算后才能得到目标变量不同，文本生成类方法可以通过直接输入提示语或提供少量示例的提问方式实现更加灵活和高效的文本变量构造。其主要应用场景包括：第一，信息识别和提取，具体可以利用生成式模型从大量非结构化文本中自动

① 该文确定了一组可能表示短视的关键词：day(-s or daily)，week(-s or-ly)，month(-s or-ly)，quarter(-s or-ly)，latter half (of the year)，short-term，short-run。

识别和提取关键信息或数据。第二，文本分类。利用生成式模型对输入文本进行分类，将分类结果直接用于生成结构化变量。Kuroki et al.(2023)利用 GPT-3.5-turbo 将公司收益电话会议文本分类为“事实”或“意见”。研究发现在电话会议文本中包含更多“意见”类陈述的企业利润更低。第三，文本总结。利用生成式语言模型的文本总结能力，从长文本或复杂文本中提炼出关键内容或文本摘要。Kim et al.(2023)利用 GPT-3.5 对公司 MD&A(管理层讨论与分析)和盈余电话会议文本分别生成文本摘要，通过计算自动摘要与原文本的长度比值得到文本膨胀度(bloat)的衡量指标，研究发现，文本膨胀度越高，公司的信息不对称程度越高、定价效率越低。第四，文本预测。利用生成式模型的理解和分析能力对文本进行分析并生成预测变量。Chen et al.(2023)使用 1996 年至 2022 年间《华尔街日报》头版的新闻标题和提示作为 ChatGPT 的输入文本，并让 GPT 从“看涨、看跌、无法判断”中选择一个预测结果，用于构造月度涨跌指标。研究发现，看涨指标能够很好地预测股票市场走势，同时与宏观经济状况具有同步性。

2. 利用文本生成方法构造虚拟代理人

生成式 NLP 方法兼具对海量信息的高效处理能力、可以被灵活赋予差异化特征，以及强大的文本生成等能力，使其在学术研究中展现出巨大潜力。其一，在经济学领域，许多研究尝试利用生成式方法构造虚拟代理人。传统经济学实验通常依赖大量真人参与，通过创造类似真实世界的实验场景来模拟和检验经济理论或市场机制的实施效果。然而，其可能存在的受试者选择性偏差、实验行为与真实行为差异可能引起的外部效度有限、实验成本较高等问题，难以大规模开展。生成式 NLP 方法能够模拟不同个体在不同经济情景中的行为，帮助经济学家探索个体决策过程、经济政策的实施效果等复杂问题。Horton(2023)利用 GPT-3 构造虚拟被试重新实施了经典的经济学实验，发现大语言模型具备理性经济人特性，可以被用于进行模拟经济学实验。Wu et al.(2023)利用深度神经网络构造 AI agents 参加一系列信任博弈的实验，研究发现经过训练后的 AI agents 做出的行为决策与人类的行为决策在本质上具有很高的相似性。这些研究初步表明了大语言模型在模拟经济学实验中的潜能。

其二，在金融学领域，生成式 NLP 方法处理多模态、多来源、大规模数据的能力，使其能够更准确地捕捉现实金融市场的行为特征；同时，问答式的运行方式赋予了其适应性和灵活性，能够提供多样化的策略分析，因此被应用于智能化金融投资策略分析。Lopez-Lira and Tang(2023)利用 ChatGPT 将新闻标题分类为好消息、坏消息和中立消息，然后用以预测当日的股票收益率，发现与 GPT-1、GPT-2 和 BERT 相比，ChatGPT 具有更好的表现。

五、经济学研究范式变革

自然语言处理技术的迭代升级，不仅革新了研究工具，更深层次地触发了经济学研究范式的系统性革新，这一变革可以归纳为三个逐渐递进、相互关联的方面。

首先，自然语言处理技术通过将海量的非结构化文本转化为可量化的特征变量，实现了研究对象的根本性拓展。传统经济学研究高度依赖财务报表、宏观经济指标等结构化数据，NLP 方法则将海量的非结构化文本数据转化为可以量化的特征指标（如词频、可读性、情绪、相似度），使之能够融入经典的计量经济模型，这极大地拓展了经济学研究的边界，使许多因关键变量难以度量而被长期搁置的重要议题（如政策不确定性、投资者情绪、年报信息含量）得以被科学度量与实证检验。

其次，随着词嵌入、主题模型及理解型大语言模型的应用，研究重点从对文本表层特征的统计分析，深化为对语义内涵的深度理解。这意味着研究已不再满足于计算词频、文本情感等浅层特征，而是致力于挖掘文本的深层语义，从而更精准地识别关键构念（如企业创新、公司文化），从而真正打开经济行为与社会互动背后的理论“黑箱”。

最后，生成式大语言模型的崛起，标志着研究逻辑从面向过去的解释性分析，转向面向未来的构建与模拟。通过构造“虚拟代理人”或进行大规模计算实验，研究者能在受控环境中大规模、低成本地模拟人类经济行为，为经济学提供了兼具内部效度与外部拓展性的全新研究路径，预示着一个“解释现实”与“模拟推演”并重的新范式正在形成。同时，生成式模型在信息提取、文本摘要方面的能力，也极大地提升了研究效率与研究深度。

六、研究结论与启示

自然语言处理技术的发展与普及，已远不止于为经济学研究增添新的工具，而是深刻触发了研究范式的系统性变革。本文通过梳理自然语言处理技术从基于统计的方法、机器学习方法、深度学习方法直至大语言模型的演进路径，并在此基础上构建了“文本分析—文本理解—文本生成”三层递进的应用框架，系统阐释了该技术如何拓展经济学研究的边界、深化其内涵并重塑其方法论基础。

本文的启示在于，面对数据形态的多元化与智能化趋势，经济学研究方法的创新必须与技术演进同步。未来研究应更主动拥抱自然语言处理等人工智

能技术，不仅将其作为数据处理的工具，更应思考如何将其深度融入理论构建、机制检验与政策模拟的全过程。同时，也需关注并规范由此带来的方法伦理、数据偏差与模型可解释性等新挑战。总之，自然语言处理正在并将持续地重塑经济学研究的疆域与逻辑，拥抱这一变革，将为理解和塑造复杂经济系统带来前所未有的可能。

参 考 文 献

[1] Bellstam, G., S. Bhagat, and J. A. Cookson, "A Text-Based Analysis of Corporate Innovation", *Management Science*, 2020, 67(7), 4004-4031.

[2] Blei, D. M., and J. D. Mcauliffe, "Supervised Topic Models", *Advances in Neural Infomration Processing*, 2007, 121-128.

[3] Blei, D. M., A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation", *Journal of Machine Learning Research*, 2003, (3), 993-1022.

[4] Bochkay, K., S. V. Brown, A. J. Leone, and J. W. Tucker, "Textual Analysis in Accounting: What's Next?", *Contemporary Accounting Research*, 2023, 40(20), 765-805.

[5] Brochet, F., M. Loumioti, and G. Serafeim, "Speaking of the Short-Term: Disclosure Horizon and Managerial Myopia", *Review of Accounting Studies*, 2015, 20(3), 1122-1163.

[6] Chen, J., G. Tang, G. Zhou, and W. Zhu, "ChatGPT, Stock Market Predictability and Links to the Macroeconomy", SSRN, 2023.

[7] Chen, L., and S. Mankad, "A Structural Topic and Sentiment-Discourse Model for Text Analysis", *Management Science*, 2024, 71(7), 5767-5787.

[8] Cohen, L., C. Malloy, and Q. Nguyen, "Lazy Prices", *The Journal of Finance*, 2020, 75(3), 1371-1415.

[9] Donovan, J., J. Jennings, K. Koharki, and J. Lee, "Measuring Credit Risk Using Qualitative Disclosure", *Review of Accounting Studies*, 2021, 26(2), 815-863.

[10] Drinkall, F., E. Rahimikia, J. B. Pierrehumbert, and S. Zohren, "Time Machine GPT", Arxiv Preprint Arxiv: 2404.18543, 2024.

[11] Dyer, T., M. Lang, and L. Stice-Lawrence, "The Evolution of 10-K Textual Disclosure: Evidence From Latent Dirichlet Allocation", *Journal of Accounting and Economics*, 2017, 64(2), 221-245.

[12] Gentzkow, M., B. Kelly, and M. Taddy, "Text as Data", *Journal of Economic Literature*, 2019, 57(3), 535-574.

[13] 韩亚楠、刘建伟、罗雄麟，"概率主题模型综述"，《计算机学报》，2021 年第 6 期第 44 卷，第 1095—1139 页。

[14] Horton, J. J., "Large Language Models as Simulated Economic Agents: What Can we Learn from Homo Silicus?", Arxiv Preprint Arxiv: 2301.07543, 2023.

[15] Huang, A., H. Wang, and Y. Yang, "FinBERT-a Large Language Model for Extracting Information From Financial Text", *Contemporary Accounting Research*, 2022, 2(40), 806-841.

[16] Huang, Z., W. Xu, and K. Yu, "Bidirectional LSTM-CRF Models for Sequence Tagging", ArXiv Preprint arXiv: 1508.01991, 2015.

[17] 姜富伟、刘雨旻、孟令超，“大语言模型、文本情绪与金融市场”，《管理世界》，2024 年第 8 期第 40 卷，第 42—64 页。

[18] 姜富伟、孟令超、唐国豪，“媒体文本情绪与股票回报预测”，《经济学》（季刊），2021 年第 4 期第 21 卷，第 1323—1344 页。

[19] Kim, A. G., M. Muhn, and V. V. Nikolaev, “From Transcripts to Insights: Uncovering Corporate Risks Using Generative AI”, Chicago Booth Research Paper No. 23-19, 2023.

[20] Kim, Y., “Convolutional Neural Networks for Sentence Classification”, Arxiv Preprint Arxiv: 1408.5882, 2014.

[21] Krishna, K., S. Ramprasad, P. Gupta, B. C. Wallace, Z. C. Lipton, and J. P. Bigham, “GenAudit: Fixing Factual Errors in Language Model Outputs with Evidence”, Arxiv Preprint Arxiv: 2402.12566, 2024.

[22] Kuroki, Y., T. Manabe, and K. Nakagawa, “Fact or Opinion? -Essential Value for Financial Results Briefing”, 2023 14th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI), 2023, 375-380.

[23] Lee, C. M. C., and Q. Zhong, “Shall We Talk? The Role of Interactive Investor Platforms in Corporate Communication”, *Journal of Accounting and Economics*, 2022, 74(2), 101524.

[24] 李春涛、闫续文、张学人，“GPT 在文本分析中的应用：一个基于 Stata 的集成命令用法介绍”，《数量经济技术经济研究》，2024 年第 5 期第 41 卷，第 197—216 页。

[25] Li, F., “Textual Analysis of Corporate Disclosures: A Survey of the Literature”, *Journal of } Literature*, 2011, 29143-165.

[26] Li, K., F. Mai, R. Shen, and X. Yan, “Measuring Corporate Culture Using Machine Learning”, *The Review of Financial Studies*, 2021, 34(7), 3265-3315.

[27] 李晓溪、杨国超、饶品贵，“交易所问询函有监管作用吗？——基于并购重组报告书的文本分析”，《经济研究》，2019 年第 5 期第 54 卷，第 181—198 页。

[28] Lopez-Lira, A., and Y. Tang, “Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models”, Arxiv Preprint Arxiv: 2304.07619, 2023.

[29] Loughran, T., and B. Mcdonald, “When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks”, *The Journal of Finance*, 2011, 66(1), 35-65.

[30] Loughran, T., and B. Mcdonald, “Textual Analysis in Accounting and Finance: A Survey”, *Journal of Accounting Research*, 2016, 54(4), 1187-1230.

[31] 马长峰、陈志娟、张顺明，“基于文本大数据分析的会计和金融研究综述”，《管理科学学报》，2020 年第 9 期第 23 卷，第 19—30 页。

[32] Obaid, K., and K. Pukthuanthong, “A Picture Is Worth a Thousand Words: Measuring Investor Sentiment by Combining Machine Learning and Photos from News”, *Journal of Financial Economics*, 2021, 144(1), 273-297.

[33] 钱贵明、阳镇、师磊，“AI 大模型产业政策体系重塑：美国经验与中国路径”，《技术经济》，2025 年第 1 期第 44 卷，第 14—27 页。

[34] 沈艳、陈赟、黄卓，“文本大数据分析在经济学和金融学中的应用：一个文献综述”，《经济学》（季刊），2019 年第 4 期第 18 卷，第 1153—1186 页。

[35] Srivastava, R. P., “A New Measure of Similarity in Textual Analysis: Vector Similarity Metric ver-

sus Cosine Similarity Metric", *Journal of Emerging Technologies in Accounting*, 2023, 20(1), 77-90.

[36] 唐国豪、姜富伟、张定胜，“金融市场文本情绪研究进展”，《经济学动态》，2016 年第 11 期，第 137—147 页。

[37] Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need", ArXiv Preprint arXiv: 1706.03762, 2017.

[38] 王靖一、黄益平，“金融科技媒体情绪的刻画与对网贷市场的影响”，《经济学》(季刊)，2018 年第 4 期第 17 卷，第 1623—1650 页。

[39] Wu, J. X., Y. D. Wu, K. Chen, and L. Hua, "Building Socially Intelligent AI Systems: Evidence from the Trust Game Using Artificial Agents with Deep Learning", *Management Science*, 2023, 69(12), 7236-7252.

[40] 姚加权、张锟澎、罗平，“金融学文本大数据挖掘方法与研究进展”，《经济学动态》，2020 年第 4 期，第 143—158 页。

[41] Zhang, Z., H. Zhang, K. Chen, Y. Guo, J. Hua, Y. Wang, and M. Zhou, "Mengzi: Towards Lightweight Yet Ingenious Pre-trained Models for Chinese", Arxiv Preprint Arxiv: 2110.06696, 2021.

Natural Language Processing and the Transformation of Economic Research Paradigms: A Framework for Text Analysis, Understanding, and Generation

YANG Guochao
(Fudan University)
YUAN Xi
(Henan University)
GONG Qiang*
(Zhongnan University of Economics and Law)

Abstract: Natural Language Processing (NLP) is a novel research methodology that relies on computers and algorithms to analyze, process, and compute human language, enabling automatic understanding and even regeneration of natural language. This paper begins by summarizing the evolution of relevant technical methods. Following the developmental trajectory of NLP approaches, it introduces the underlying principles of key models, including

* Corresponding Author: GONG Qiang, Wenlan School of Business, Zhongnan University of Economics and Law, Wuhan, Hubei 430073, China; Tel: 86-27-88386678; E-mail: qianggong@zuel.edu.cn.

Word2Vec, BERT, and large language models. Furthermore, the paper outlines an application framework for NLP methods from the three perspectives of analyzing, understanding, and generating text. The advancement of NLP technology is driving a threefold transition in economic research paradigms: from reliance on structured data to unstructured text, from quantification of surface features to deep understanding, and from explaining the past to simulating the future.

Keywords: natural language processing; machine learning; large language model

JEL Classification: B41, C40, G00