文本分类算法及其应用场景研究综述

刘晓明 李丞正旭 吴少聪 张宇辰 白红艳
程泽华 陈 卓 李永峰 兰 钰 沈 超
(西安交通大学电子与信息学部 西安 710049)

摘 要 随着大数据时代的到来,互联网中的文本信息迎来了井喷式的增长.文本分类作为自然语言处理中最重要的技术之一,其广泛应用于多个领域,如情感分析、新闻分类、自然语言推理、主题标记、抽取式问答、虚假内容检测等.从传统机器学习分类方法理论的深入到深度学习分类方法探索的兴起,相关研究模型与思路也在不断演变,各类新的方法、数据集和评价指标层出不穷,丰富了文本分类领域的研究,取得了卓越的理论成就和应用效果.尽管如此,新技术不断发展和业务应用场景不断丰富,同时,也为文本分类研究带来了许多新的问题与挑战,如数据约束场景中不均衡数据的文本表征学习、小样本场景下的文本分类等.针对当前研究难题与挑战,本文对文本分类方法进行了系统性调研,并对当前方法在实际应用场景中面临的技术挑战和未来的研究方向进行了综合探讨.具体而言,本文主要综述了七部分内容,分别是:(1)对文本分类技术的相关基础知识进行了全面介绍,包括文本分类的常见符号定义、计算范式和文本预处理技术;(2)对基于传统机器学习的文本分类方法进行了详细总结;同时,为了方便读者针对不同的应用场景选择合适的分类模型,本文对不同分类器擅长处理的文本分类难题及方法优劣进行了总结;(3)对基于新兴深度学习的文本分类方法进行了周详梳理,根据领域内代表性技术的核心思想进行分类,在此基础上对不同类别下的主要方法进行描述,同时对其技术的优劣进行了总结;(4)为了方便读者对文本分类模型的有效性进行验证,针对文本分类技术应用最为广泛的七大场景,本文对相关数据集进行了系统性的总结;(5)本文对不同任务目标下的常用的模型评价方法进行详尽介绍,以便对模型性能进行合理的定量评估;(6)基于上述内容,本文对典型应用场景中不同种类文本分类算法进行了性能总结对比;(7)本文分别从数据约束与模型计算两个层面对当前文本分类技术所面临的挑战和未来的重要研究方向进行了总结.本文通过梳理文本分类研究发展脉络,对涉及的代表性技术进行了详细总结和对比分析,有效填补了文本分类领域前沿技术的应用综述空白.

关键词 文本分类;机器学习;深度学习;评价指标;数据约束
中图法分类号 TP391 **DOI 号** 10.11897/SP.J.1016.2024.01244

A Survey of Text Classification Algorithms and Application Scenarios

LIU Xiao-Ming LI Cheng-Zheng-Xu WU Shao-Cong ZHANG Yu-Chen BAI Hong-Yan
CHENG Ze-Hua CHEN Zhuo LI Yong-Feng LAN Yu SHEN Chao
(*Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an* 710049)

Abstract With the advent of the era of big data, text information on the internet has ushered in a blowout growth. As one of the most important technologies in natural language processing,

收稿日期:2024-02-08;在线发布日期:2024-02-09.本课题得到国家重点研发计划(2020YFB1406900)、国家自然科学基金(62272371,61902308,U21B2018,62103323,62161160337,61822309,61773310)、博士后创新人才支持计划基金(BX20190275,BX20200270)、博士后面上基金(2019M663723,2021M692565)、中央高校基础科研经费(xzy012024144)、陕西省重点产业创新计划项目(2021ZDLGY01-02)资助.**刘晓明**(通信作者),博士,副教授,中国计算机学会(CCF)会员,主要研究方向为社交网络对抗、时空异常事件检测、机器学习方法及其应用.E-mail:xm.liu@xjtu.edu.cn.**李丞正旭**,硕士研究生,主要研究方向为强化学习、文本智能计算.**吴少聪**,硕士研究生,主要研究方向为推荐系统、知识图谱挖掘、文本智能计算.**张宇辰**,博士研究生,主要研究方向为社交网络对抗、虚假新闻检测.**白红艳**,硕士研究生,主要研究方向为群体发现、图神经网络应用、文本智能计算.**程泽华**,硕士研究生,主要研究方向为机器生成文本检测、文本智能计算.**陈 卓**,硕士研究生,主要研究方向为图神经网络应用、文本智能计算.**李永峰**,博士研究生,主要研究方向为异常检测.**兰 钰**,博士,助理教授,主要研究方向为系统优化、机器学习及其应用.**沈 超**,博士,教授,长江学者,中国计算机学会(CCF)杰出会员,主要研究领域为人工智能可信与安全、信息物理系统控制与安全、智能软件安全与测试.

text classification has a wide range of applications, such as sentiment analysis, news categorization, natural language inference, topic labeling, extractive question answer and fake news detection, etc. From the deepening of traditional machine learning methods to the rising of deep learning methods, related research of text classification models and ideas are constantly evolving, and various new methods, data sets, and evaluation indicators emerge in an endless stream, enriching the research in the field of text classification and achieving excellent theoretical achievements and application effects. Nevertheless, with the rapid development of advanced new technologies, the rich and diverse business application scenarios have also introduced many complex new technical challenges to this field, such as text representation learning with unbalanced data, text classification under few-shot learning scenarios, and so on. In response to the above research challenges and problems, this paper conducts an overall survey of text classification methods, and comprehensively discusses the technical challenges faced by current methods and future research directions. More specifically, this paper mainly consists of seven parts, which are (1) Introducing the relevant basic knowledge of text classification technology, including the definition of common symbols, computational paradigms and text preprocessing techniques, and so on. (2) Summarizing the text classification methods based on traditional machine learning. At the same time, in order to facilitate readers to select the appropriate models for different application scenarios, this paper summarizes the advantages and disadvantages of different classifiers, i. e., what kind of text classification problems they are good at dealing with. (3) Sorting out the text classification methods based on the emerging deep learning carefully, which are classified according to the key ideas of representative technologies in the field. Then the main methods under different categories are described, in which their advantages and disadvantages are summarized thoroughly. (4) In order to facilitate readers to verify the validity of the text classification models, this paper systematically summarizes the relevant datasets for the seven most widely used scenarios of text classification technology. (5) This paper introduces the commonly used model evaluation methods under different task objectives in detail, so as to quantitatively and reasonably evaluate the text classification model performance. (6) Based on the above, this paper summarizes and compares the performance of different types of text classification algorithms in typical application scenarios. (7) Summarizing the challenges faced by existing text classification technology and the important research directions in the future from two aspects, i. e., data limitation and model computation performance. By sorting out the development of text classification research, this paper provides a detailed summary and comparative analysis of representative technologies involved in the development of text classification research which effectively addresses the gap in the application overview of innovative technologies in the field of text classification and offers a comprehensive reference for researchers to quickly get started on related issues.

Keywords　text classification; machine learning; deep learning; evaluation metrics; limited data

1　引　言

文本分类是指在给定分类体系下，通过特定模型计算，为输入文本指定预定义标签的过程，是自然语言处理(Natural Language Processing，NLP)中应用最广泛、也是最重要的领域之一. 目前，文本分类已在现实中多个领域得到了应用，包括情感分析[1-2]、主题标记[3-4]、问答任务[5-6]、新闻分类[7-8]等. 这些领域技术所应用的服务场景与我们的日常生活息息相关，例如智能客服[9-10]、商业智能[11-12]和舆情评估[13-14]等. 由此可见，文本分类作为一项基础性技术显得尤为重要. 与此同时，随着学术界、工业界对该问题愈发重视，其领域内的关键技术也在不断演变.

最初，人们使用词匹配法来进行文本分类[15-16]，

通过文本中是否出现与类名相同的词来对文本类别进行判断,但这种简单机械的方法显然无法带来良好的分类结果.为了改善词匹配法有效性不足的问题,人们后续采用知识工程方法[17],借助于专业人员的帮助,为每个类别定义大量的推理规则,进而通过规则匹配的方式来判断文本类别.这类方法通过引入人为判断,准确度相较词匹配法大幅提高.但是,近年来随着大数据、云计算等技术的迅速发展,互联网信息迎来了爆炸式的增长,人工手动设计规则处理和分类大量文本数据的方式面临巨大的挑战.此外,该类方法的分类准确率极易受到人为因素的影响,例如专业人员的知识储备和精力因素等.在这之后,人们尝试使用机器学习方法来实现文本分类自动化[18].该类方法通过模拟人类对大量同类文本的特征进行学习,作为今后分类的依据,由此能够得到更加可靠且客观的分类结果.

总体来说,从 1960 年到 2010 年左右,基于传统机器学习的文本分类方法占据主导地位.该类方法在分类过程中大多遵循以下三个步骤:文本预处理、特征提取(例如,词袋模型[19]和 TF-IDF[20]等)以及分类计算(使用朴素贝叶斯(Naïve Bayes,NB)[21]、K 近邻算法(K-Nearest Neighbor,KNN)[22]、支持向量机(Support Vector Machines,SVM)[23]等).尽管这些方法在分类准确性和稳定性方面表现良好,但它们在实际应用中存在明显的限制.例如,它们高度依赖于耗时且成本高昂的特征工程.此外,由于强烈依赖于领域知识,该类方法在新的分类任务中的可扩展性和有效性受限.不仅如此,这些方法往往忽略了文本数据中的序列信息、上下文信息和单词本身的语义信息,这与人类对句子的理解过程不符.同时,由于语言知识本身具有的笼统性、复杂性和歧义性等特点,如何对文本内蕴含的语义信息进行有效挖掘面临着严峻的挑战[24].

近年来,深度学习方法在自然语言处理、图像识别和语音识别等领域取得了突破性进展[25],受到了工业界和学术界广泛关注,这也为文本分类的研究带来了新的机遇.一方面,与前文所论述的传统方法相比,深度学习方法避免了人工设计规则和特征的过程,可通过学习一种深层次非线性网络结构,自动从样本中挖掘出文本中的本质特征,能够捕获文本数据的深层次语义表征信息;另一方面,深度学习模型通过对多模态数据进行特征学习,从而将多个模态的信息共同映射到联合向量空间,能够获得数据的统一表征[26].在此基础上,模型能够有效融合多模态特征来对文本类型进行综合判断,缓解从单一数据进行判断所面临的语义歧义、信息匮乏等问题,使得识别和分类更加准确、可靠.此外,鉴于深度学习网络的高度可扩展性,众多深度学习网络架构应运而生,从不同角度对文本数据进行建模,以应对文本分类任务中语义理解、特征提取、数据不平衡等一系列挑战.因此,2010 年以后,文本分类方法研究重心已逐渐从传统机器学习方法迁移到新兴深度学习方法.这类方法的核心模块是一个表征学习模型,该模型能够自动挖掘潜在特征,并将文本数据映射到蕴含语义信息的低维连续特征向量,因此不需要传统机器学习方法中的人工设计特征.例如,2013 年 Google 公司研究人员提出了 word2vec 模型[27],其核心思想为利用单词所在上下文信息将非结构化、不可计算的文本数据转换为结构化、可计算的低维稠密向量;2018 年华盛顿大学研究人员提出了 ELMo 模型[28],基于多层双向 LSTM 网络对单词的复杂特征(例如句法和语义特征)和在语境中的变化进行建模,实现单词的复杂表示学习.此外,近几年随着 Transformer 模型[29]的兴起,BERT[30]、GPT-2[31]、GPT-3[32]等一系列大规模预训练语言模型陆续被提出,该类方法无需人工标签,可以从海量的语料中学习到通用的语言表示,并显著提升下游的任务,将 NLP 领域的研究提升到了一个新的阶段.其中,文本分类作为 NLP 领域的核心任务,大规模预训练语言模型在跨域、低资源、少样本等复杂场景下展现出了卓越的性能表现.

综上所述,近年来文本分类领域的研究呈现出丰富而杂乱的局面.尽管已经有一些优秀的英文综述论文[33-36]对文本分类技术进行了总结和归纳,但随着预训练语言模型[37]、迁移学习[38]、Prompt 模型[39]等新技术的涌现,这些技术的突破不断增强了文本分类领域的发展,同时也催生了一些新的挑战和应用场景.因此,现有的综述缺乏一定的时效性.此外,本文也注意到在数据集方面,文本分类相关的数据集也日益增多,但缺少系统性的梳理和整理.本文对主流数据集进行系统性的梳理,为研究者提供更全面和准确的数据支持.此外,现有的综述大多从单一或部分技术路线进行综述,目前尚缺少较为细致、系统、全面的文本分类方法梳理工作.例如,苏金树等人[40]和 Minaee 等人[34]只对机器学习方法或深度学习方法进行综述,而庞亮等人[41]和薛春香等人[42]等则仅聚焦于某个领域的文本分类方法.因此,本文沿着该领域的发展脉络,对文本分类算法从

传统机器学习模型到新兴深度学习模型的研究和应用进行了系统性综述，并对比分析了其在不同场景下的优劣势.此综述便于研究人员快速了解文本分类领域的发展历程和最新技术，并展望了未来的研究方向和挑战，为后续的研究工作提供基础和灵感.

整体论文组织框架如图 1 所示，其中第 2 节对文本分类的相关基础知识进行介绍，包括本文常用的符号定义、文本分类方法范式和文本分类基础知识；第 3 节梳理了传统的文本分类算法，主要包含常见的机器学习分类研究工作；第 4 节综述了深度学习的主要方法，基本涵盖了所有主流的深度学习网络架构；第 5 节整理了已有相关研究数据集，包含文本分类算法的主要应用领域；第 6 节介绍了算法的评价方法，分为单标签和多标签分类任务等；第 7 节对不同类型文本分类算法在常见的应用场景下进行了综合性能对比；第 8 节分别从数据约束、模型计算两个层面总结了目前文本分类面临的主要挑战，并对未来研究趋势进行展望；第 9 节总结了全文内容.

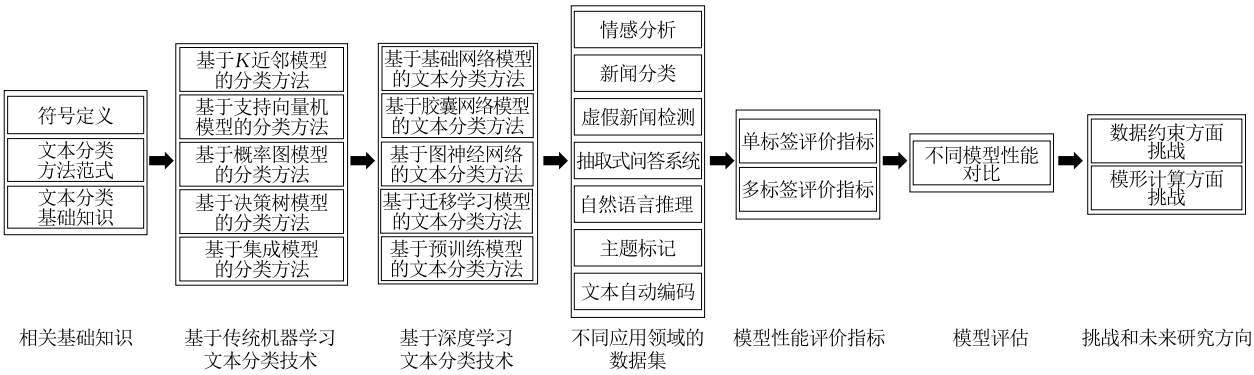

图 1　文章组织框架

2　相关基础知识

本节将给出文中常见的符号定义、文本分类方法范式和文本分类基础知识.

2.1　符号定义

表 1 罗列出了文中常用符号及其代表含义.文本数据集可以用 $D=\{S_1,S_2,\cdots,S_N\}$ 表示，由一系列长短不一的文本构成，N 表示数据集所包含的文本数目；$S=\{w_1,w_2,\cdots,w_n\}$ 表示单个文本数据，其中 w_i 表示文本中第 i 个单词，n 表示文本数据所包含的单词数目；$Y=\{y_1,y_2,\cdots,y_m\}$ 表示分类任务中的类别信息，其中 m 表示类别数目.此外，针对其它信息，本文利用 $\boldsymbol{w}_i\in R^{1\times d}$ 表示单词 w_i 的嵌入向量；$\boldsymbol{M}\in R^{n\times d}$ 表示由文本 S 中所有单词的特征向量所构成的特征矩阵，其中每一行表示一个单词的特征向量.

表 1　符号定义

符号	表示含义
σ	激活函数
D	数据集/文档集合
D_i	类别 y_i 下样本集合
$S\in D$	单个文本数据
$w_i\in S$	文本中第 i 个单词
n	文本数据大小
Y	分类标签集合
$y_i\in Y$	第 i 个分类标签
m	分类标签数量
$\boldsymbol{w}_i\in R^{1\times d}$	单词 w_i 的特征向量
$\boldsymbol{M}\in R^{n\times d}$	文本 S 的特征矩阵
Dic	词汇字典数据
V	词汇表
Θ	模型参数
K	KNN 最近邻数量
ξ_i	SVM 松弛变量

2.2　文本分类方法范式

图 2 表示在传统机器学习方法和新兴深度学习方法的基础上，文本分类系统计算的基本范式.对于输入文本数据，传统机器学习方法主要包含 3 个步骤：首先，需要进行文本预处理，达到文本中的噪声信息消除、文档分词等效果；其次，进行特征提取以获得良好的样本特征（部分工作将以上两个步骤统称为特征工程）；进而，使用经过训练的基于统计学的机器学习算法对样本进行分类.与传统机器学习方法不同，新兴的深度学习方法首先通过学习文本的潜在特征嵌入表示；随后，利用设计的非线性变换神经网络模型将特征映射到系统输出，以此完成文本分类.基于上述结果，根据不同的任务目标采用相应的评价指标对方法的分类效果进行评估，最终达到情感分析、新闻分类等目的.

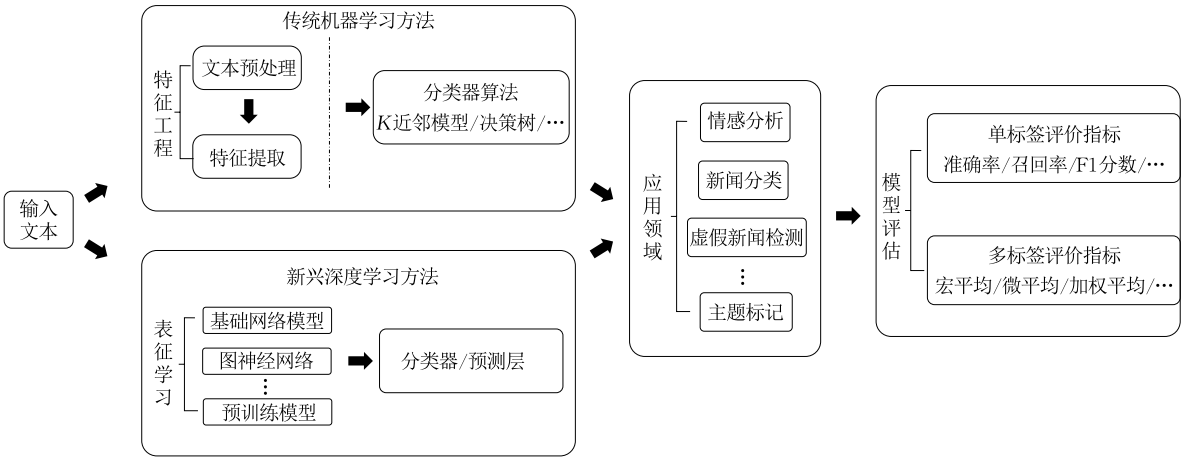

图 2 文本分类流程图(传统机器学习方法包含文本预处理、特征提取和分类器算法 3 个步骤;深度学习方法通过将特征工程集成到表征学习中,能够自动化地提取特征)

2.3 文本分类基础知识

虽然文本分类方法在传统机器学习和新兴深度学习两个方面的基本范式具有差异,但是在发展过程中不同分类技术的作用机理具有一定的逻辑性和相似性.为方便读者进行理解,本小节将针对不同领域的文本分类相关基础知识展开简要叙述.

传统机器学习方法在解决文本分类问题时由特征工程和分类器两个模块组成,其中特征工程包含文本预处理和特征提取两个部分内容.通常情况下,文本或文档都是非结构化数据.然而,在作为分类器输入时,非结构化文本序列必须转换为可处理的结构化形式,文本预处理和特征提取是该过程的关键步骤.大多数情况下,文本数据都包含许多不必要的单词信息,例如无特殊意义的停用词、拼写错误单词、俚语等.在许多算法中,这些噪声信息和不必要的特征会对算法性能产生负面影响.因此,在文本分类之前对数据进行预处理是有必要的.相关技术例如:分词技术、停用词处理、拼写检查等.分词是一项 NLP 基础任务,由于原始文本数据粒度太大,承载的信息量多,面临难以复用等问题.因此,需要通过分词技术切断上下文耦合关系,降低词序影响,将文本数据分解为字词、短语或其它能够表达完整含义的较小单位[43-44],方便后续的进一步处理和分析.针对中文文本,由于词与词之间没有空格等标志来指示分割边界,这使得分词过程具有一定挑战性.目前主流方法采用基于统计的分词方式,通过动态规划查找最大概率路径,以找出基于词频的最佳切分组合.关于停用词处理技术,由于通常情况下文本数据中会包含许多出现频率非常高,但并不具有重要实际意义的词,例如数量词“a”、“another”、“一”、“另一”,以及一些介词或者连词“about”、“above”、“否则”、“然而”等,这些词被统一归类为停用词.针对这类停用词,最直接的做法就是在不牺牲句子含义的情况下,安全地将它们从文本数据中删除,以此来节省存储空间和提高算法性能[45].拼写检查是可选的文本预处理步骤.拼写错误最常出现在社交媒体相关的文本数据集中(例如新浪微博、Twitter、Instagram 等),由于人们的操作失误导致误输入.目前,在 NLP 领域已经有很多方法能够解决拼写问题[46],结合深度学习技术,研究人员可以使用语言模型或者目前流行的预训练模型来对拼写问题进行纠正[47].

在完成文本预处理步骤后,需要对规范后的数据进行特征提取,将非结构化文本序列转换为结构化特征空间数字向量,才能够作为机器学习分类器的输入进行后续的分类计算.在传统机器学习分类方法中,主流的特征提取技术包括词袋模型、词频-逆文档频率等.词袋模型(Bag of Words,BoW)[19]是常用的文档简化表达模型,该模型利用由单词集合构成的词汇表来表示每个文档,词袋中的每个单词互相独立,并在计算过程中保留了单词的统计特征,示例如下:

“*I like apples, and I like oranges, too.*”

对于上述句子,单词集合即词袋为

“*I*” “*like*” “*apples*” “*and*” “*oranges*” “*too*”

句子的词袋特征可表示为

$$[2,2,1,1,1,1]$$

词汇表中的每个单词均通过独热编码(one-hot)进行表示,对句子中所有单词的编码表示进行求和即得到最终的词袋特征.虽然 BoW 模型实现起来非常

简单，但是它在特征构造中忽略了单词之间的顺序关系，且无法反映出文档中不同单词之间的关联关系，而这些往往是表达句子主题的关键信息.

词频-逆文档频率算法(Term Frequency-Inverse Document Frequency，TF-IDF)[20]是一种基于词语统计加权的向量化表示方法. 其中，TF(Term Frequency)表示单词在文档中出现的次数(频数)，简称词频. 其形式化表示如下：

$$TF(w,D)=\frac{\text{单词 } w \text{ 在文章 } D \text{ 中出现的次数}}{\text{文章 } D \text{ 的总词数}} \tag{1}$$

IDF(Inverse Document Frequency)表示单词的逆文本频率指数，即当一个单词在文档中出现越频繁，对应的 IDF 数值越小，反之，IDF 值越高. IDF 与 TF-IDF 的形式化表示如下：

$$IDF(w)=\log\frac{\text{语料库中文章总数}}{\text{包含单词 } w \text{ 的文章总数}+1} \tag{2}$$

$$TF\text{-}IDF(w,D)=TF(w,D)\times IDF(w) \tag{3}$$

TF-IDF 值的大小虽然能够反映文本中单词的重要性，但每个单词都作为索引独立呈现，没有考虑单词之间的关联关系，同样也忽略了单词之间的顺序关系.

近年来随着复杂模型的发展，出现了新的特征提取技术来弥补传统方法的缺点，例如词嵌入模型. 词嵌入作为一种特征学习技术，其目标是将词汇表中的每个单词映射到一个特定维度实数向量，包括 Word2Vec[48]、GloVe[49]、FastText[50]等模型.

2013 年，Mikolov 等人[48]提出了“word to vector”的表示方法来作为改进的词嵌入架构. Word2Vec 模型常用的两种神经网络结构包括 CBOW(Continuous Bag-of-Words Model)模式和 Skip-gram 模型. 其中，CBOW 模型核心思想是根据上下文来预测中心目标单词，计算某个词出现的概率，通过极大似然估计获取最终每个单词的嵌入表示；而 Skip-gram 模型则相反，是根据某个中心词分别计算出该词前后出现的某几个词，即上下文单词出现的各个概率. 模型一次输入的字数通常取决于局部滑动窗口大小的设置(根据统计学结果，常用大小为 4～5 个字). Word2Vec 模型通过将单词映射到向量空间后，相似的词在向量空间中更为接近，可以较好地发现文本语料库中的序列关系以及单词之间的相似性. 无独有偶，2014 年斯坦福大学提出的 GloVe 词嵌入方法[49]同样地将单词映射到向量空间之中，使词向量能够包含尽可能多的语义和语法信息. 模型的目标函数定义如下：

$$f(\boldsymbol{w}_i,\boldsymbol{w}_j,\widetilde{\boldsymbol{w}}_k)=\frac{P_{ik}}{P_{jk}} \tag{4}$$

其中，$\boldsymbol{w}_i$ 表示单词 w_i 的词向量，P_{ik} 和 P_{jk} 表示单词 w_k 在单词 w_i 和单词 w_j 的上下文中出现的概率. GloVe 模型的词向量构造核心思想是：首先基于语料库构造词的共现矩阵，然后通过对共现矩阵进行矩阵分解，可以获取每个单词的低维表征向量. GloVe 模型既使用了语料库的全局统计(overall statistics)特征，也使用了局部的上下文(local context)特征. Word2Vec 与 GloVe 模型都是通过学习一个不同的向量来表示语料库中的单词. 虽然这些方法考虑了单词之间的相似性或语料库的全局特征以及部分上下文特征，但是却忽略了单词的内部结构. 2016 年，Facebook AI Research 提出了 FastText 框架[50]解决了这一问题. 该模型使用了字符级别的 n-grams 来表示一个单词，从而改进形态丰富语言的向量表示. 例如，对于单词“apples”，当 $n=3$ 时，该单词表示为以下字符三元组集合：

$$[\text{“}app\text{”}\ \text{“}ppl\text{”}\ \text{“}ple\text{”}\ \text{“}les\text{”}\ \text{“}es\text{”}]$$

语境化词表征通常基于 context2vec[51]词嵌入方法实现，其中 context2vec 方法通过长短时记忆神经网络(Long Short-Term Memory Networks，LSTM)来对语句中单词蕴藏的信息进行学习. 2018 年华盛顿大学研究人员[28]利用这一思想提出了深层语境化词表征技术，构建了 ELMo 模型. 该方法对单词的复杂特征与其在语境中的变化进行建模，其生成的单词向量通过双向语言模型学习得到，其中 BiLM 由前向和后向语言模型组成. 在那之后，Transformer 模型[29]开始兴起，BERT[30]、GPT-2[31]等一系列大规模预训练语言模型陆续被提出，可以从海量的语料中提取通用的文本特征，显著提升了下游的任务. 文本分类作为 NLP 领域内的核心任务，其相关研究进入了一个新阶段.

3　文本分类机器学习模型

如图 2 所示，传统的机器学习文本分类方法的工作范式是在经过停用词处理、词干提取等文本预处理[52]后，利用词袋模型[19]、TF-IDF[20]等方法进行特征提取，进而将其作为分类器的输入来进行预测. 常用的传统文本分类方法主要包括：K 近邻算法、支持向量机、概率图、决策树等. 本节将分别对不同传统机器学习文本分类方法进行阐述与总结，对比分析各类方法的应用场景及其优劣势，各类方法的基本信息如表 2 所示.

表 2 传统文本分类方法对比总结

传统文本分类方法	方法优点	方法缺点	应用场景
基于 K 近邻模型	方法简单直观，并且准确率较高，可处理非线性文本	计算复杂度较高，难以处理大规模任务，分类结果受噪声文本和不均衡文本影响较大	情感分析[57]，新闻分类[59]
基于支持向量机模型	能够捕获非线性特征，具有较强泛化能力	计算复杂度高且易受缺失及噪声数据影响，对核函数选择敏感	新闻分类[63-64]
基于概率图模型	算法简单且分类效率稳定，对缺失数据的敏感程度低	模型描述能力有限，受不均衡数据分布影响较大	情感分析[65]，新闻分类[66]
基于决策树模型	基于决策树的方法易于解释和理解，分类效率高	对数据扰动较为敏感，容易出现过拟合问题	新闻分类[67]，情感分析[68-69]
基于集成模型	可通过级联不同的模型来提高预测性能和分类效率，易于实现，泛化能力比单个模型较强	计算复杂度高，多个模型级联将导致可解释性下降	新闻分类[70-71]，虚假新闻检测[72]

3.1 基于 K 近邻模型的分类方法

K 近邻算法（K-Nearest Neighbors，KNN）最早由 Cover 等人[53]提出，被广泛应用于诸多研究领域. 其核心思想是给定一个测试数据 x，首先计算 x 与各个训练数据之间的距离并按照距离的递增关系进行排序，随后选取与 x 距离最小的 K 个数据，进一步确定前 K 个数据中各类别的出现频率，将出现频率最高的类别作为测试数据 x 的分类结果.

当 KNN 用于文本分类时[54]，通常使用文本预处理和特征提取方法将给定测试文档 S 表示为结构化特征向量，使用欧几里得距离、余弦距离等方法[55]计算出 S 与训练集中所有文档的距离，从而选出最近的 K 个邻居作为 S 的最近邻. 接着利用距离衡量 S 与每个近邻文档的相似性，根据这 K 个近邻文档的类别对测试文档 S 的候选类别进行加权评分，最终将得分最高的候选类别分配给测试文档 S. KNN 的决策规则如下：

$$f(S)=\arg\max_j Score(S,y_j)=\sum_{S_i\in kNN} sim(S,S_i)h(S_i,y_j) \tag{5}$$

其中，$f(S)$表示测试文档 S 通过模型计算获得的文本标签，$Score(S,y_j)$表示文档 S 与对应候选类别 y_j 的分数，$sim(S,S_i)$表示测试文档 S 与训练集中文档 S_i之间的相似性，$h(S_i,y_j)\in\{0,1\}$是训练集中文档 S_i是否属于类别 y_j 的二分类值（$h=1$ 表示 S_i属于类别 y_j，反之则 $h=0$）.

KNN 通过与其他方法结合，可解决数据约束场景下的文本分类问题. 2020 年，Liu 等人[56]针对多类少样本学习问题，将可学习 KNN 分类器和 MLP 与注意力模块相结合，以产生对粗粒度类别和细粒度类别的可靠预测，在文中针对小样本学习问题提出的基准数据集上提高了 1.9%的准确率. 2021 年，Isnain 等人[57]利用 KNN 分类方法针对印尼语推特数据进行情感分析，当 $K=10$ 时，获得了最优分类效果，准确率达到了 84.7%. 2021 年，Kamalloo 等人[58]提出了 MiniMax-KNN 方法作为一种半监督的数据增强技术，提高了基于 KNN 方法的数据增强性能. 2022 年，Shi 等人[59]提出了 KNN-Prompt 方法，利用 KNN 检索增强语言模型，提高小样本数据下的分类效率. 实验表明，该方法在 AG News[60]、YahooA[61]等数据集下，与 GPT-2 模型增强之前对比，性能平均提高 13.4%，在 SST-2[62]数据集上达到了 84.2%的准确率.

基于 K 近邻模型的传统分类器简单、有效且直观的特点，被广泛应用于情感分析、新闻分类等任务. 然而，这一类方法严重依赖于处理文本时所使用的文本预处理和特征提取方法，模型的性能提升往往得益于特征提取方法的进步发展，缺少对所提取特征的深层次信息挖掘与利用，难以对不同类别数据实现细粒度区分. 基于 KNN 的文本分类方法，因此，对于现实应用场景中存在的数据不平衡问题、数据分布不一致问题、类别边界模糊问题等都缺少有效的解决手段. 此外，这类方法具有较高的时空复杂度，对部署环境的硬件运行速度与数据存储空间都有较高的要求，难以解决大规模的文本分类任务.

3.2 基于支持向量机模型的分类方法

支持向量机（Support Vector Machine，SVM）由 Cortes 等人[73]提出，其作为有效的分类算法适用于众多分类任务. 该模型具有的较高泛化能力，使其适用于文本等高维数据分类任务.

当 SVM 用于文本分类任务时[74]，其核心思想是将数据集中的每个文本表示为一个向量，通过超平面隔离使得支持向量距离最大化，实现文本数据分类. 令$(\boldsymbol{x}_1,y_1),\cdots,(\boldsymbol{x}_l,y_l)$表示一组数据，其中 $\boldsymbol{x}_i\in R^N$ 表示文本向量，$y_i\in\{-1,+1\}$表示正负样本. 考虑

形式为 $\mathrm{sgn}((\boldsymbol{w}\cdot\boldsymbol{x})+b)$的决策函数，其中$(\boldsymbol{w}\cdot\boldsymbol{x})$表示 $\boldsymbol{w}$ 和 $\boldsymbol{x}$ 的内积，SVM 模型的目的是找到一个具有式(6)性质的决策函数 $f_{w,b}$.

$$y_i((\boldsymbol{w}\cdot\boldsymbol{x})+b)\geq 1,\ i=1,\cdots,l \tag{6}$$

然而，在许多实际文本分类应用场景下，SVM 分类超平面并不存在. 为解决这一问题可引入松弛变量 $\xi_i\geq 0$，满足以下条件：

$$y_i((\boldsymbol{w}\cdot\boldsymbol{x})+b)\geq 1-\xi_i,\ i=1,\cdots,l \tag{7}$$

在文本分类任务中，支持向量机方法经常被用于提高有限样本条件下的文本分类精度. 2012 年，Wan 等人[75]提出了 SVM-NN 方法，通过 SVM 来改善 KNN 中的参数 K 较小时文本分类精度降低的问题. 2013 年，Garla 等人[76]针对电子医疗记录数据标注样本不足的问题，采用半监督的 Laplacian SVM 方法实现了对未标记语料的有效利用. 2018 年，Goudjil 等人[77]利用 SVM 提出了一种新的文本分类主动学习方法 AL-MSVM，在降低数据标记需求的同时实现了分类精度的提升. 此外，支持向量机方法还被证明能够有效应对新闻分类问题. 2016 年，Dadgar 等人[63]基于 TF-IDF 和 SVM，提出了一种新的新闻分类文本挖掘方法，在 BBC[78]和 20 Newsgroups[79]数据集上分别实现了 97.8%和 94.9%的分类精准度. 2019 年，Londo 等人[64]在针对印尼新闻文章进行分类方法的研究中，通过对比不同的传统文本分类方法，发现 SVM 取得了较高的性能，$F1$ 分数达到了 0.93.

支持向量机模型能够捕捉文本数据中的非线性关系，提高文本分类的泛化能力，对于改善有限样本条件下的模型分类精度具有良好效果. 然而，这一方法针对非线性分类问题缺少通用解决方案，其核函数构造较为复杂，使该方法缺乏足够的可解释性. 同时该方法对缺失数据、噪声数据较为敏感，因此难以解决较复杂的文本分类问题.

3.3　基于概率图模型的分类方法

概率图模型(Probabilistic Graphical Model，PGM)是用于计算在一组变量 $\boldsymbol{X}$ 上定义的联合概率分布，由一个包含 $\boldsymbol{X}$ 中每个变量 X_i(节点表示)及变量之间因果关系(有向边表示)的图表示. 基于 PGM 的一些分类方法主要包括贝叶斯网络(Bayesian Networks，BNs)、马尔可夫网络(Markov Networks，MNs)、隐马尔可夫模型(Hidden Markov Models，HMM)等. 其中，朴素贝叶斯(Naïve Bayes，NB)是使用最广的基于 PGM 的模型之一，应用于多个研究领域，例如文本分类[80]、细胞关联推断[81]等. 以下将以朴素贝叶斯分类器为例，介绍基于 PGM 的文本分类方法核心思想.

朴素贝叶斯方法的 PGM 结构如图 3 所示，其主要使用先验概率来计算后验概率，如式(8)所示：

$$P(y|X_1,X_2,\cdots,X_n)=\frac{P(y)\prod_{i=1}^{n}P(X_i|y)}{\prod_{i=1}^{n}P(X_i)} \tag{8}$$

其中，y 表示类别，$\boldsymbol{X}=[X_1,X_2,\cdots,X_n]$表示样本 $\boldsymbol{X}$ 的 n 维特征向量.

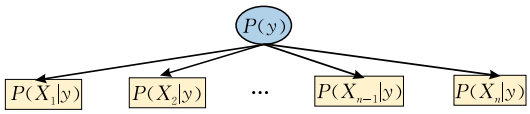

图 3　朴素贝叶斯 PGM 结构

当 NB 分类器用于文本分类时[80]，文档 $S=\{w_1,w_2,\cdots,w_L\}$的类别 $y\in Y=\{y_1,y_2,\cdots,y_m\}$可以由其特征通过条件概率计算获得.

根据最大后验概率(Maximum A Posteriori probability，MAP)决策规则，样本文档S的类别y由式(9)决定：

$$y=\arg\max_{y\in Y}P(y|S)=\arg\max_{y\in Y}P(y)P(S|y) \tag{9}$$

基于概率图模型的文本分类方法在数据约束的场景下获得了较好的性能. 2016 年，Ebrahimi 等人[82]基于铰链损失马尔可夫随机场(Hinge-Loss Markov Random Fields，HL-MRFs)提出了一种针对 Twitter 数据的弱监督文本分类器. 2018 年，Kang 等人[65]提出了一种新的基于文本隐马尔可夫模型(TextHMMs)的情感分析方法，该方法在训练文本时使用词序列而不是预定义的情感词典. 2020 年，Mukherjee 等人[66]通过贝叶斯网络将不确定性考量引入自训练过程中，从而改善预训练微调(fine tuning)模型在少量数据标签情况下的文本分类性能. 该方法在 SST-2[62,83]、AG News[60]等数据集上进行了实验，平均准确率为 91.0%.

基于概率图模型的分类方法模型结构简单，计算时间复杂度较低，且对缺失数据敏感程度不高，在情感分析、新闻分类等任务上取得了有效运用. 但是这类方法对于数据独立性有着较强的假设，且概率图构建对数据的先验概率计算具有较强依赖性，因此基于概率图模型的文本分类算法性能受数据稀缺性，尤其是数据不均衡分布的影响，容易在分类学习过程中形成偏差，难以对文本数据进行多层次、细粒度深入分析.

3.4 基于决策树模型的分类方法

决策树(Decision Trees,DT)模型[84]是一种有监督的自顶向下递归构造的树结构学习方法,其主要目标是构建一个模型,该模型可通过从训练数据学习决策规则预测目标变量的类别.在构造决策树后,模型通常可通过剪枝来减少噪声及过拟合对于分类结果的影响.

基于决策树模型的文本分类方法同样在数据约束场景下发挥着作用.2008 年,Vens 等人[85]使用层级多标签分类(Hierarchical Multi-label Classification,HMC)决策树对功能基因组学进行研究.他们的研究表明 HMC 决策树可以学习到标签之间的依赖关系,同时具备较好的可解释性.2016 年,Bahassine 等人[67]提出了 ImpCHI 方法作为一种新的特征选择技术,并应用于阿拉伯文本数据的决策树模型分类任务中.2019 年,Taloba 等人[68]通过结合决策树和遗传算法,提出了一种新的混合机器学习技术 GADT,用于检测垃圾邮件,达到了降低数据噪声影响的效果.实验表明,与 KNN、SVM 等其它传统模型相比,GADT 获得了较好的性能增益,达到了 95.5%的准确率.2021 年,Yuvaraj 等人[69]提出了一种新的深度决策树分类器用于识别网络欺凌文本,通过消除噪声和其它背景信息,提取选定的特征并进行分类,降低数据过拟合对结果的影响.实验表明,深度决策树分类器达到了 93.6%的准确率.

基于决策树模型的分类方法简明直接的特点,其易于理解和解释.同时,作为一种非常快速的学习和预测算法,它能够提供较高的文本分类效率,适用于处理大规模文本数据的分类场景.但是这类方法容易发生过拟合问题,对数据的泛化能力不足;同时也对数据中的扰动较为敏感,数据中微小的分布变化可能会导致决策树发生较大的性能改变.

3.5 基于集成模型的分类方法

集成学习(Ensemble Learning,EL)通过构建多个分类器,并根据一定的规则对各个分类器的结果进行聚合,来完成文本分类任务.集成模型的学习方式主要包括 Boosting(如梯度提升决策树(Gradient Boosting Decision Tree,GBDT)[86])和 Bagging(如随机森林[87]以及 Stacking[88]).

针对不同数据约束条件下的文本分类任务,基于集成模型的分类方法同样取得了较好的效果.2011 年,Shi 等人[70]提出了一种基于粗糙集和集成学习的文本分类半监督算法,降低了未标记数据集对于文本分类任务的影响.2016 年,Kilinç[71]评估了集成学习模型对土耳其语数据集进行文本分类的影响,结果表明集成学习模型在很大程度上提高了基础分类器的分类精度.例如,J48 基础分类算法准确率为 77.1%,当使用 Bagging 和 Boosting 集成模型时,准确率分别提高了 4.6%和 8.4%.2019 年,Al-Ash 等人[72]提出了一种针对印尼虚假新闻数据的集成学习方法,一定程度上解决了在给定数据集上的数据不平衡问题.实验表明,使用随机森林分类方法的 $F1$ 分数为 0.98,而非集成分类算法的多项式朴素贝叶斯和支持向量机分类方法的 $F1$ 分数分别为 0.43 和 0.74.2020 年,Zeng 等人[89]提出了一种基于集成学习的短文本合格性标准分类器,该方法集成了 BERT、ERNIE 等预训练模型,实验准确率为 84.6%,性能平均高于单个模型 2.4%,减少了类不平衡的影响.

集成模型应用于文本分类时的泛化能力和效率优于单个模型,能够利用模型的多样性和鲁棒性来解决文本数据存在的高维度、稀疏性和不均衡性等问题.但是计算复杂度相对更高,同时多个模型级联将导致方法整体可解释性下降.

总体而言,传统机器学习方法与早期基于规则的方法[17]相比,能够适应于诸如情感分析、新闻分类、邮件检测、文本识别等多种不同的文本分类任务,同时表现出更高的分类准确率和模型效率.但是,受限于模型本身复杂度的限制,传统机器学习方法缺乏对文本数据深入分析的能力,往往忽略了文本数据中的自然语序结构或全局上下文信息,也难以有效学习文本数据包含的深层语义信息.因此,已有的多种基于传统机器学习方法的文本分类模型都普遍存在泛化能力不足、易受数据干扰、输出结果可解释性差等问题.在最近的各项研究中,传统机器学习方法较少被视作为文本分类模型的主干,而更多是搭配各类深度学习模型使用.通过"深度学习模型提取特征、传统机器学习完善推理"的方式,研究者试图挖掘传统机器学习方法在数据约束条件下存在的推理优势,配合能够深层次分析文本数据的深度学习模型解决各类问题.由此,文本分类方法逐渐由传统机器学习的浅层分类方法发展成为深度学习模型,各类文本分类深度学习模型逐渐进入广大研究者视野.

4 文本分类深度学习模型

近年来,深度学习技术已经成为大数据与人工

智能领域的研究热点[90]. 与传统机器学习模型相比，深度学习模型通过对特征间的高阶交互关系进行学习，能够得到表达能力更强的数据分布表征，挖掘出更多数据中潜在的模式. 基于这一特点，深度学习技术被广泛应用于各个领域，例如图像识别、机器翻译、语音识别、文本分类等，并取得了突破性进展. 如图 4 所示，本节沿着深度学习技术在文本分类领域应用的发展脉络，对不同阶段所涉及的代表性模型进行了详细介绍和讨论.

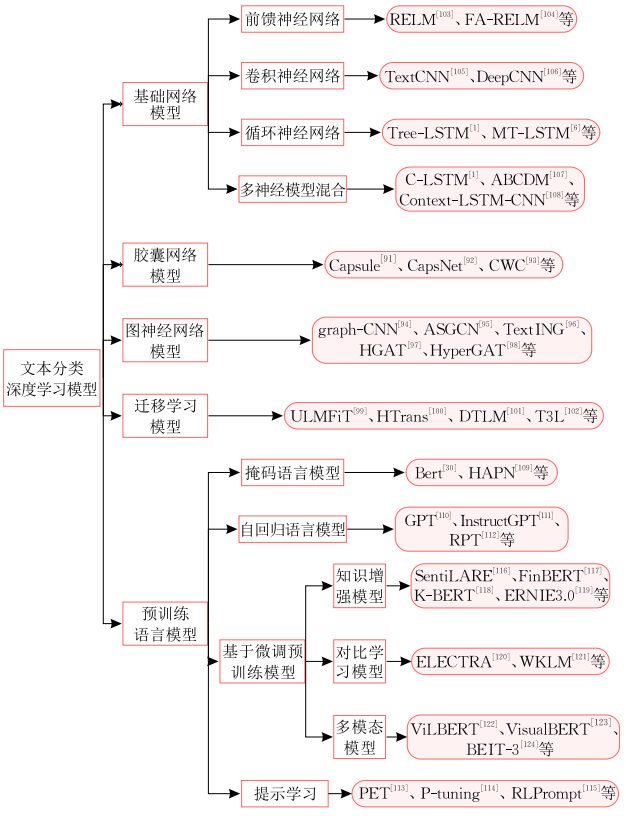

图 4　基于深度学习的文本分类方法及其代表性工作

4.1　基于基础网络模型的分类方法

如何从文本中提取和表示有效特征，一直以来都是文本分类任务的一大挑战. 对比传统机器学习方法通常依赖繁琐的手动特征工程，深度学习方法能够自动挖掘潜在特征模式，降低特征工程带来的资源消耗成本. 基础的深度学习网络模型主要包括前馈神经网络[125]、循环神经网络[126]、卷积神经网络[127]等，研究人员针对不同模型特点提取文本的多维特征，并将其应用在文本分类领域，取得了显著效果.

前馈神经网络(Feed-Forward Neural Networks，FNN)是用于文本表示学习的最基础的深度学习模型之一[103-104]. FNN 主要包括五层结构：输入层、嵌入层、堆叠层、神经网络层、输出层. 2016 年，Joulin 等人[50]提出了一个简单有效的文本分类模型，名为 FastText. 其核心思想是在输入层除了使用连续词袋特征之外，还利用单词的字符级别 n-grams 向量作为额外特征. 该过程同时考虑了单词内部的形态和单词之间的顺序特征，并将它们进行叠加平均来得到更加丰富的文本特征表示. 并且，由于在过程中不同单词的 n-grams 特征相互之间可以进行共享，因此相较于其它方法，该模型能够学习到更好的低频词特征. 另外，对于预测过程中遇到的词库之外的单词(Out-Of-Vocabulary，OOV)，该模型仍然可以通过叠加相应的字符级 n-grams 特征向量来构建它们的词向量. 虽然全连接神经网络(FNN)是许多深度学习方法的基础，但它在处理文本时的表达能力受到结构上的限制. FNN 在顺序信息的捕捉上主要依赖上下文窗口大小，因此对于长单词或文本，它难以有效地捕获远距离的顺序依赖. 此外，由于 FNN 结构的简单性，它在利用词序、单词位置、上下文和时序信息等关键文本特征方面表现不足. 这些限制导致 FNN 在理解语义复杂的文本内容方面面临挑战.

为了挖掘更为丰富的文本语义信息，应对长文本特征提取场景下的上下文依赖问题，研究人员尝试使用循环神经网络(Recurrent Neural Network，RNN)对文本中的时序特征进行建模. 具体而言，自然语言文本能够被视为一种典型的带有时间顺序的序列信息，例如对于文本 $S=\{w_1,w_2,\cdots,w_n\}$，其中单词 $w_i\in S$ 即为其在 t_i 时刻的信息. 对于这类信息的处理，RNN 是一种常用的有效手段[128]. 但是基础的 RNN 模型在处理时间步数较大的信息，如长文本语义信息时，非常容易出现梯度消失的问题，从而导致模型对于长距离信息的遗忘. 为此，在文本分类领域所用的多为 RNN 相关变体，这些变体通过引入相应模块对信息传输进行控制，来实现对信息的记忆功能，如门控循环神经网络(Gated Recurrent Neural Network)[129]、长短期记忆(Long Short-Term Memory，LSTM)网络[130]等. 其中 LSTM 通过引入了记忆细胞(memory cell)和三种门：输入门(input gate)、遗忘门(forget gate)和输出门(output gate)来实现对信息的控制. 具体 LSTM 的状态转移方程如下：

$$\boldsymbol{I}_t=\sigma(\boldsymbol{X}_t\boldsymbol{W}_{xi}+\boldsymbol{H}_{t-1}\boldsymbol{W}_{hi}+\boldsymbol{b}_i) \tag{10}$$

$$\boldsymbol{F}_t=\sigma(\boldsymbol{X}_t\boldsymbol{W}_{xf}+\boldsymbol{H}_{t-1}\boldsymbol{W}_{hf}+\boldsymbol{b}_f) \tag{11}$$

$$\boldsymbol{O}_t=\sigma(\boldsymbol{X}_t\boldsymbol{W}_{xO}+\boldsymbol{H}_{t-1}\boldsymbol{W}_{hO}+\boldsymbol{b}_O) \tag{12}$$

$$\widetilde{\boldsymbol{C}}_t=\tanh(\boldsymbol{X}_t\boldsymbol{W}_{xc}+\boldsymbol{H}_{t-1}\boldsymbol{W}_{hc}+\boldsymbol{b}_c) \tag{13}$$

$$\boldsymbol{C}_t=\boldsymbol{F}_t\odot\boldsymbol{C}_{t-1}+\boldsymbol{I}_t\odot\widetilde{\boldsymbol{C}}_t \tag{14}$$

$$\boldsymbol{H}_t=\boldsymbol{O}_t\odot\tanh(\boldsymbol{C}_t) \tag{15}$$

其中，$\boldsymbol{X}_t$代表在时间步 t 的输入，$\boldsymbol{H}_t$代表 t 时间步的隐藏状态；$\boldsymbol{I}_t$、$\boldsymbol{F}_t$、$\boldsymbol{O}_t$对应着时间步 t 时刻的输入门、遗忘门和输出门，$\boldsymbol{W}_{xi}$与 $\boldsymbol{b}_i$ 分别为权重参数和偏置参数；$\boldsymbol{C}_t$与$\widetilde{\boldsymbol{C}}_t$分别代表着时间步 t 时刻的记忆细胞和候选记忆细胞.

在此基础上，2015 年 Tai 等人[1]提出一种树结构的 LSTM 模型，并将其用到文本分类领域，取得了优秀的性能表现. 该模型的基本思想是传统的 LSTM 虽然可以对自然语言文本中的序列信息进行分析，但是难以分析自然语言文本中的树结构信息（如根据句法生成的树状依赖关系），故提出了一种树结构的 LSTM 用以处理文本中的树结构信息. 相较于传统 LSTM，树结构的 LSTM 不只从上一时刻的数据获取信息，还从子节点获取信息. 借助树结构的 LSTM，该模型在情感分类任务上取得了优于传统方法的性能. 2015 年，Liu 等人[6]提出了多时间尺度的 LSTM 结构（Multi-Timescale LSTM，MT-LSTM），通过将标准 LSTM 的隐藏状态分成几组，每组在不同时间段激活来实现对长文档的建模. 由此，MT-LSTM 在文本分类任务上达到了优于基线方法的性能. 鉴于现有方法大多依赖预先指定的文本结构，2018 年 Zhang 等人[131]将强化学习（Reinforcement Learning，RL）[132]方法与 LSTM 结合，提出了一种自动探索和优化文本结构的方法 HS-LSTM，用以学习文本表示. 实验结果表明，HS-LSTM 无需明确的结构设计，可通过识别重要词或任务相关结构来学习文本表示，从而提升模型文本分类性能.

循环神经网络在处理长文本特征提取方面取得了进展，但对于某些特定的文本分类任务，如情感分析，捕捉文本的局部特征（例如短语或关键词）可能比长期依赖更为关键. 因此除时序信息外，研究人员还尝试使用卷积神经网络（Convolutional Neural Network，CNN）[133]对空间信息进行建模，挖掘文本中的局部特征. 为了尝试将 CNN 应用于文本分类任务，2014 年 Kim 等人[105]提出了一种无偏卷积神经网络模型，即 TextCNN. 该模型的核心思想是利用多个不同大小的滤波器来提取文字中的关键信息，从而能够更好地捕捉文本中的局部相关性和确定相关判别短语. 具体对于利用一个滤波器提取特征过程的形式化表示如下：

$$\boldsymbol{w}_{1:|S|}=\boldsymbol{w}_1\oplus\boldsymbol{w}_2\oplus\cdots\oplus\boldsymbol{w}_{|S|} \tag{16}$$

$$c_i=f(\boldsymbol{W}\cdot\boldsymbol{w}_{i:i+h-1}+b) \tag{17}$$

$$\boldsymbol{c}=[c_1,c_2,\cdots,c_{|S|-h+1}] \tag{18}$$

$$\hat{c}=\text{Max-Pooling}(\boldsymbol{c}) \tag{19}$$

其中，对于输入层，$\boldsymbol{w}_i\in\mathcal{R}^{1\times d}$表示句子 S 中第 i 个单词的 d 维词向量，$\oplus$是连接操作，$\boldsymbol{w}_{1:|S|}$是句子的特征矩阵.

对于卷积层，与一般 CNN 不同的是该模型的滤波器为 $W\in\mathcal{R}^{k\times d}$，其宽度与词向量维度一致，窗口大小为 k，c_i为从单词 $\boldsymbol{w}_{i:i+h-1}$窗口生成的特征值，$\boldsymbol{c}\in\mathcal{R}^{|S|-h+1}$是利用该滤波器卷积得到的特征图. 对于池化层，该模型通过在特征图上执行最大池化操作，以此提取特征图中最重要的特征 $\hat{c}$. 如图 5 所示，该模型使用多个具有不同窗口大小的滤波器来获取多个特征，这些特征进行拼接得到倒数第二层的句子表示，并传递给后续的分类器，计算得到类别标签上的概率分布. 后续实验表明，即使是固定住词向量不参与训练，仅仅利用一个带有少量超参调整的 CNN 进行预测，仍然能够在多个基准测试中取得很好的分类结果.

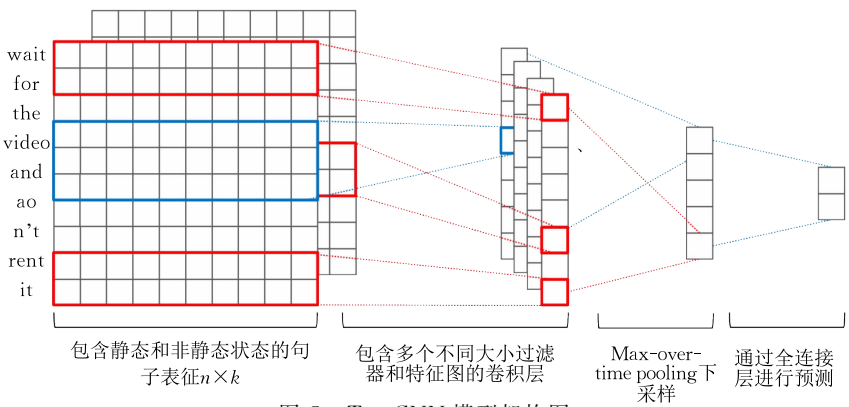

图 5 TextCNN 模型架构图

尽管如此，由于卷积核尺寸通常不会很大，因此 TextCNN 在捕获长距离特征时效果不理想，导致在长文本分类场景下效果不佳. 此外，该方法仅是 CNN 在文本分类领域的初步尝试，后续仍有较大的提升空间. 基于文本的最小嵌入单元，可以将嵌入方法分为字符级（character-level embeddings）、单词级（word-

level embeddings)和句子级嵌入(sentence-level embeddings).其中字符级嵌入可以解决词汇表之外的单词嵌入问题;单词级嵌入可以学习单词之间的语法和语义信息;句子级嵌入可以捕捉句子之间的潜在关联.基于上述特点,2015 年 Zhang 等人[134]提出了字符级 CNN 模型,核心思想为仅利用字符级信息进行文本分类过程.该模型将输入文本表示为固定长度的字符向量矩阵,其中每个字符用相应的独热编码进行表示,随后通过由六个具有池化操作的卷积层和三个全连接层构成的深度 CNN 模型,计算得到最终的预测标签.作者从实验的角度证明了基于卷积神经网络的文本分类无需语言的语法和语义结构的知识也能够取得良好的效果,同时分析了该模型在不同文本分类任务上的适用性.但是字符级别的文本分类方法由于长度限制,不适用于长文本分类场景,以及忽略文本中的语义信息的缺陷可能会限制模型的性能.2017 年,Nguyen 等人[106]提出了一种基于词典的深度学习方法,即 DeepCNN 模型.该模型核心思想为通过构建语义规则处理非必要的子句,以及利用深度卷积神经网络来生成能够捕获单词的形态和形状信息的字符嵌入,用于增强单词级嵌入表示信息,进而提高分类准确率.

考虑到在复杂文本分类场景中不同神经网络模型各自的优势,也有研究人员尝试将多种神经网络模型混合,从不同角度共同对文本的深层语义进行建模,提升整体模型的效果.例如,2015 年 Zhou 等人[135]通过将 CNN 与 RNN 结合构建了 C-LSTM 模型.该模型首先通过数个卷积层从单个语句中提取文本局域特征,之后将提取的特征输入 LSTM 进行进一步处理,相较于传统的 LSTM 方法,特征提取过程由纯粹词向量序列变成了经过卷积的抽象含义序列.在 TREC[136]数据集上的实验表明,C-LSTM 模型相比于最优基线基线模型 Bi-LSTM 获得了 1.6%的性能增益,获得了 94.6%的准确率.2018 年,Petrak 等人[108]针对大多数方法忽略待分类的句子与相邻的句子所构成的上下文之间的语境影响这一问题,提出了一种新的分类方法 Context-LSTM-CNN.该模型核心思想是在利用 Bi-LSTM 和 CNN 捕获目标句子中的远程依赖关系和局部相关性的同时,通过 FOFE(Fixed Size Ordinally Forgetting)方法[137]建模目标句子的上下文语境信息.考虑到在复杂文本分类场景下,文本数据长度不均匀对模型处理能力提出的额外挑战,2021 年 Basiri 等人[107]提出了 ABCDM 模型,该模型结合了双向 GRU (Gated Recurrent Unit)、LSTM 以及 CNN,共同对文本中的长期依赖和局部特征进行建模.在训练过程中,首先使用预训练的 GloVe 词向量模型对输入文本 $S=\{w_1,w_2,\cdots,w_n\}$ 进行嵌入,得到文本对应的嵌入矩阵 $\boldsymbol{M}\in\boldsymbol{R}^{n\times d}$,其中 d 为向量维度大小,$\boldsymbol{w}_t\in\boldsymbol{M}$ 表示单词 w_t 的词向量表示.接着,将 $\boldsymbol{M}$ 并行作为 Bi-LSTM 和 Bi-GRU 的输入,利用这两个模型对任意长度的序列进行处理,提取其前向和后向的长依赖关系,具体计算过程如下:

$$\overrightarrow{\boldsymbol{h}}_{t_{\mathrm{LSTM}}}=\overrightarrow{\mathrm{LSTM}}(\boldsymbol{w}_t),\ t\in[1,n] \tag{20}$$

$$\overleftarrow{\boldsymbol{h}}_{t_{\mathrm{LSTM}}}=\overleftarrow{\mathrm{LSTM}}(\boldsymbol{w}_t),\ t\in[n,1] \tag{21}$$

$$\overrightarrow{\boldsymbol{h}}_{t_{\mathrm{GRU}}}=\overrightarrow{\mathrm{GRU}}(\boldsymbol{w}_t),\ t\in[1,n] \tag{22}$$

$$\overleftarrow{\boldsymbol{h}}_{t_{\mathrm{GRU}}}=\overleftarrow{\mathrm{GRU}}(\boldsymbol{w}_t),\ t\in[n,1] \tag{23}$$

因此,对于每个单词,可以通过连接前向和后向上下文来获得表征,具体如下所示:

$$\boldsymbol{h}_{t_{\mathrm{LSTM}}}=[\overrightarrow{\boldsymbol{h}}_{t_{\mathrm{LSTM}}},\overleftarrow{\boldsymbol{h}}_{t_{\mathrm{LSTM}}}] \tag{24}$$

$$\boldsymbol{h}_{t_{\mathrm{GRU}}}=[\overrightarrow{\boldsymbol{h}}_{t_{\mathrm{GRU}}},\overleftarrow{\boldsymbol{h}}_{t_{\mathrm{GRU}}}] \tag{25}$$

随后,为了建模文本中单词的重要性,该模型采用了注意力机制提取关键词表征信息,从而得到文本表征,其形式化如下:

$$\boldsymbol{u}_{t_{\mathrm{LSTM}}}=\tanh(\boldsymbol{W}_{\mathrm{LSTM}}\boldsymbol{h}_{t_{\mathrm{LSTM}}}+\boldsymbol{b}_{\mathrm{LSTM}}) \tag{26}$$

$$\boldsymbol{u}_{t_{\mathrm{GRU}}}=\tanh(\boldsymbol{W}_{\mathrm{GRU}}\boldsymbol{h}_{t_{\mathrm{GRU}}}+\boldsymbol{b}_{\mathrm{GRU}}) \tag{27}$$

$$\alpha_{t_{\mathrm{LSTM}}}=\frac{\exp(\boldsymbol{u}_{t_{\mathrm{LSTM}}}^{\mathrm{T}}\boldsymbol{u}_{w_{\mathrm{LSTM}}})}{\sum_t\exp(\boldsymbol{u}_{t_{\mathrm{LSTM}}}^{\mathrm{T}}\boldsymbol{u}_{w_{\mathrm{LSTM}}})} \tag{28}$$

$$\alpha_{t_{\mathrm{GRU}}}=\frac{\exp(\boldsymbol{u}_{t_{\mathrm{GRU}}}^{\mathrm{T}}\boldsymbol{u}_{w_{\mathrm{GRU}}})}{\sum_t\exp(\boldsymbol{u}_{t_{\mathrm{GRU}}}^{\mathrm{T}}\boldsymbol{u}_{w_{\mathrm{GRU}}})} \tag{29}$$

$$\boldsymbol{M}_{\mathrm{LSTM}}=\sum_t\alpha_{t_{\mathrm{LSTM}}}\boldsymbol{h}_{t_{\mathrm{LSTM}}} \tag{30}$$

$$\boldsymbol{M}_{\mathrm{GRU}}=\sum_t\alpha_{t_{\mathrm{GRU}}}\boldsymbol{h}_{t_{\mathrm{GRU}}} \tag{31}$$

其中 $\boldsymbol{W}$ 和 $\boldsymbol{b}$ 分别表示网络的参数矩阵和偏置向量,$\boldsymbol{u}_t$ 是 $\boldsymbol{h}_t$ 的隐藏表征,$\boldsymbol{u}_w$ 是随机初始化得到的上下文向量,并在训练过程中进行学习;单词 w_t 的注意力权重 α_t 是通过计算 $\boldsymbol{u}_t$ 和 $\boldsymbol{u}_w$ 之间的相似度得出的.

基于注意力权重向量 $\boldsymbol{\alpha}_t$,词特征向量被聚合到新的文本特征矩阵 $\boldsymbol{M}_{\mathrm{LSTM}}$ 和 $\boldsymbol{M}_{\mathrm{GRU}}$ 中.然后将 $\boldsymbol{M}_{\mathrm{LSTM}}$ 和 $\boldsymbol{M}_{\mathrm{GRU}}$ 作为 CNN 的输入,挖掘其中的局部 n-gram 特征,并通过一系列操作(如池化和批标准化)得到隐藏表征 $\boldsymbol{h}_p$.最后,将 $\boldsymbol{h}_p$ 作为全连接层的输入,用于预测分类结果.实验结果表明,ABCDM 在不同长度的文本分类任务上都取得了显著的效果提升.其中在 Amazon-Review[138]相关数据集上,ABCDM 取得了超过 90.0%的预测准确率.考虑到层级多标签文本分类任务复杂的标签层次结构,2021 年 Chen 等人[139]

将文本与标签之间的交互描述为一个语义匹配问题，结合 Bi-GRU 与 CNN 提出了一种层次感知的标签语义匹配网络 HiMatch 模型. HiMatch 通过不同的网络组合分别提取文本语义和标签语义，并利用文本和标签之间的语义相关性分配标签. 实验表明 HiMatch 可以有效挖掘文本-标签的语义匹配关系，并在 RCV1-2[140] 数据集上获得了平均 80.7% 的预测准确率.

总体而言，基础网络模型相较于机器学习模型，在考虑词序信息和对文本进行深层语义建模方面取得了更好的效果. 然而，基础网络模型本身也存在着模型过于简单、只能建模特定信息等问题. 例如循环神经网络对文本中的时序特征进行建模，善于处理拥有丰富上下文关系的长文本分类任务，而卷积神经网络通过挖掘文本的 n-gram 信息，更善于捕捉文本的局部特征. 虽然部分研究人员尝试将多种神经网络模型结合起来，以缓解这些单一模型的局限性，但这种方法往往需要大量的调参和实验来找到最优的模型结合方式. 因此，在未来的工作中，研究人员开始探索更高效的模型架构，考虑挖掘不同种类的潜藏信息和模型架构本身等角度寻求改进方向.

4.2 基于胶囊网络模型的分类方法

胶囊网络(Capsule Network，CN)这一概念于 2017 年由 Sabour 等人[141] 最先在计算机视觉领域提出，其最初目的是解决 CNN 池化层带来的信息丢失问题，并给出包含更精准空间信息的输出. 在胶囊网络中，胶囊(capsule)结构取代了传统的神经元(neural)结构，每个胶囊的输入和输出均为向量. 且在反向传播之外，网络同时使用动态路由(dynamic routing)机制进行网络训练. 在 Sabour 等人给出的网络架构中，激活的胶囊对应着输入图像中的特定实体，而胶囊中的神经元对应着该实体的各种属性. 胶囊给出的输出向量的长度对应着该实体的存在概率. 为此，胶囊网络中存在如下的压缩机制：

$$\boldsymbol{v}_j=\frac{\|\boldsymbol{s}_j\|^2}{1+\|\boldsymbol{s}_j\|^2}\frac{\boldsymbol{s}_j}{\|\boldsymbol{s}_j\|} \tag{32}$$

其中 $\boldsymbol{v}_j$ 代表胶囊 j 的输出向量，$\boldsymbol{s}_j$ 代表该胶囊的输入向量，$\boldsymbol{u}_i$ 是低层胶囊的输出.

除第一层胶囊外，每个胶囊的输入均为低层胶囊输出向量的加权和，即

$$\boldsymbol{s}_j=\sum_i \boldsymbol{c}_{ij}\hat{\boldsymbol{u}}_{j|i},\quad \hat{\boldsymbol{u}}_{j|i}=\boldsymbol{W}_{ij}\boldsymbol{u}_i \tag{33}$$

其中，$\boldsymbol{W}_{ij}$ 为权重矩阵，$\boldsymbol{c}_{ij}$ 则由动态路由机制产生. 图 6 展示了单层胶囊网络的运算方式.

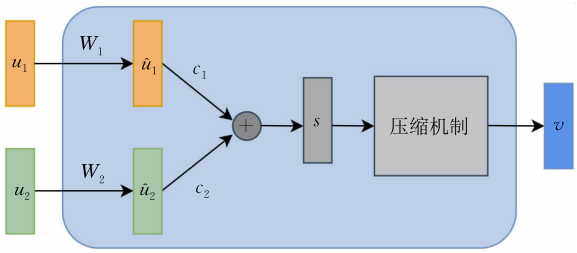

图 6 单层胶囊网络的运算方式

胶囊网络的有效性在计算机视觉任务中已得到证明[142-143]. 鉴于传统神经网络在处理文本分类任务时，可能会过度关注局部特征(例如停用词和与特定类别无关的词汇)，这种偏向导致其在分析复杂文本时性能下降. 近几年，研究人员开始探索使用胶囊网络来编码文本数据中局部与整体之间的内在空间关系，以此来减轻局部特征噪声对模型的不良影响. 2018 年，Zhao 等人[91] 将胶囊网络应用于文本分类任务，并提出了三种策略来改进动态路由机制. 实验表明，该研究所提出的三种策略减少了噪音胶囊的干扰，并使模型性能得到了提升. 通过将胶囊网络和其他文本分类模型(如 CNN、LSTM)等进行了对比，结果显示胶囊网络的性能不弱于已有方法，且在多标签迁移任务中的准确率高达 95.4%，远高于 CNN 和 LSTM. 2020 年，Kim 等人[92] 同样将胶囊网络应用于文本分类任务，在基准数据集上取得了优秀的结果. 该项研究发现由于实体的相对位置在文本和图像中有不同的影响，直接将图像的动态路由机制应用于文本数据可能并不适用. 因此，作者使用静态路由(static routing)对动态路由进行了替换，实验证明了这一改进在降低计算复杂度的同时提升了模型的准确性. 同样针对路由机制，2023 年 Zhao 等人[144] 提出了分层胶囊网络路由框架，以应对层级多标签文本分类中的误差传播问题. 该模型在 DBPedia[145] 主题标记数据集上达到了 96.1% 的准确率. 此外，研究实验表明了胶囊网络相比其他方案，在较少参数的情况下便可达到较好的分类效果.

目前，胶囊网络在文本分类领域的研究尚处于进展阶段，其相较于其他模型具有以下优点：(1) 胶囊网络中胶囊与实体的对应关系使得它可以学习和保存实体的各种属性，而这些参数可以便捷地应用到其他任务上；(2) 胶囊网络相较传统 CNN 网络可以有效建模复杂文本的空间信息，因此在面对多标签、长文本等场景下的文本分类任务时展现出更好的效果. 不过由于特殊的路由机制，目前胶囊网络也存在着如训练较慢，计算成本较高等缺点. 因此，针对文本分类的特定场景改进路由机制是一个重要的

研究方向.

4.3　基于图神经网络模型的分类方法

近年来图神经网络(Graph Neural Network, GNN)[146-147]已经在许多领域取得了优越的性能表现,如语言翻译[148]、关系分类[149]等.如前所述,自然语言文本中包含着多种词序相关的外部信息,例如上下文信息、时序和局部空间特征等.除此之外,文本内部也存在着丰富的图结构信息,例如句法和语义依存树[150]等.图神经网络能够对此类信息进行有效的挖掘,可以显著提升模型的分类效果.

对于给定的图结构数据,基于特定的信息传播方法[151],图中节点表示通过与其邻居表示的聚合学习来迭代更新.具体而言,GNN 中的第 k 层如下所示:

$$\boldsymbol{a}_t^k=\mathrm{AGGREGATE}^k(\boldsymbol{h}_s^{k-1}:s\in\mathcal{N}(t),\boldsymbol{h}_t^{k-1}) \tag{34}$$

$$\boldsymbol{h}_t^k=\mathrm{COMBINE}^k(\boldsymbol{h}_t^{k-1},\boldsymbol{a}_t^k) \tag{35}$$

其中,$\mathcal{N}(t)$ 表示与节点 t 直接相连的邻居集合,$\boldsymbol{a}_t^k$ 表示来自邻居的聚合信息,$\boldsymbol{h}_t^k$ 表示节点 t 在第 k 层的表征信息.

此外,该框架可以灵活地插入不同的 AGGREGATE 和 COMBINE 函数,从而产生各种 GNN 架构.

在文本分类领域,GNN 在处理不同粒度的文本分类问题上表现出色.由于图结构的灵活性,GNN 可以轻松地适应不同粒度的文本数据,从单词级别到整个文档级别.这种多尺度的适应性使得 GNN 成为了处理各种文本分类任务的强大工具,无论是短文本的情感分类还是长文本的主题分类.2004 年,Mihalcea 等人[152]所提出的 TextRank,是最早的基于图结构的文本分类模型之一.该方法将自然语言文本表示为图结构 $G=(\mathcal{V},\mathcal{E})$,其中 $\mathcal{V}$ 表示节点集合,$\mathcal{E}$ 表示节点之间的边集合.在实际应用场景中,节点可以表示各种文本单元,如单词、搭配词和整个句子等;边可以用于表示节点之间不同类型的关系,如语义关系、上下文重叠关系等.

在文档这一粒度,2018 年 Peng 等人[94]提出了一种 graph-CNN 的深度学习模型,该模型的核心思想是首先将文本转换为词语图,然后使用图卷积运算对词语图进行信息表征,最终通过分类器获得文档标签.研究通过实验表明,文本的词语图表示具有捕捉非连续和长距离语义的优势,而 CNN 模型则具有学习不同级别语义的优势.同样针对文档粒度的分类,2019 年 Yao 等人[7]将单词和文档同时作为节点,构建了异构的单词-文本图来表示文本数据,从而实现对文档的分类.具体而言,所构造的图中单词-文本间边的权重由词频生成,单词-单词间边的权重由其语义相关性生成,如图 7 所示.该研究使用图卷积神经网络(Graph Convolutional Network, GCN)[153]对数据进行处理以实现文本分类.其核心思想是借助 GCN,经由共同的邻居节点实现文档之间的信息传递,从而达到通过已标记文档信息学习其余单词和文档嵌入表示的目的.最终,与最优基线 Graph-CNN-C[154]相比,该方法在多个基准数据集上获得了平均 1.2%的性能增益,其平均准确率达到了 84.5%.

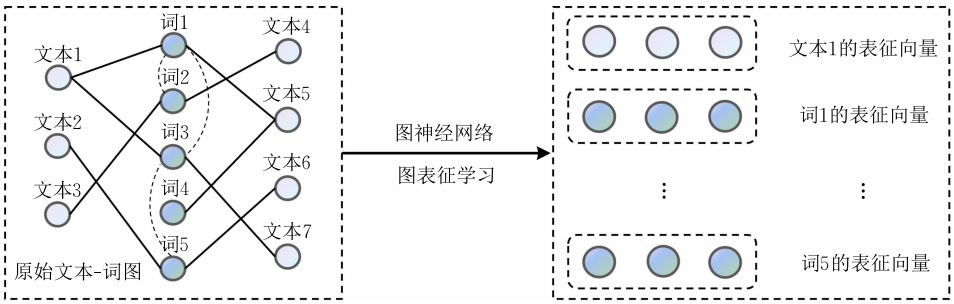

图 7　图神经网络文本表征模型

针对语句粒度的分类,2019 年 Zhang 等人[95]通过语句中依赖树(Dependency Tree, DT)结构完成文本图结构的构建,进而通过 GNN 实现对于语句的分类计算.实验结果表明 GNN 对于词的长距离依赖关系的利用提升了分类模型的性能.2020 年,Zhang 等人[96]为每篇文档构建单独的图,并使用 GNN 学习其中的单词表示.该研究提出的方法可以有效为文档中的未知单词生成嵌入表示,并且在基准数据集上表现出了优越的性能.

此外,一些学者尝试将其他方法与 GNN 方法相结合来解决文本分类任务中存在的挑战.针对 GNN 在短文本表征学习过程中存在的标记数据有限性和稀疏性等挑战,2019 年 Hu 等人[97]通过将注意力机制与 GNN 相结合,提出了异构图注意力网络(Heterogeneous Graph Attention Networks, HGAT).HGAT 不仅可以集成不同类型附加信息的框架来对短文本进行建模,而且利用了双重注意力机制来学习相邻节点的重要性以及不同节点类型

的重要性.实验表明 HGAT 在短文本分类上体现出来明显的优势,并在多个基准数据集上有更好的性能表现.但是,大多数基于 GNN 的文本分类方法主要关注单词之间的二元关系.然而,在自然语言中,文字之间的关系并不仅限于二元关系,还存在高阶关系.例如,对于文本"eat humble pie",它的意思一般是"承认自己错了".若模型只建模(eat-pie)之间的二元关系,则可能会将单词"pie"误解为"烤盘",进而误解整个习语的实际含义.为解决该问题,2020 年 Ding 等人[98]提出了 HyperGAT 模型,该模型通过文档级超图(hypergraphs)来对文本进行建模.与二元关系相比,超图建模能够更好地捕获单词的高阶异构上下文信息,例如顺序和语义.此外,HyperGAT 还对传统的 GNN 模型进行改进,提出了一个基于双重注意机制的超图模型,以支持文本超图的表征学习.通过这些改进,HyperGAT 不仅可以提高模型的计算效率,而且可以更准确地对文本进行建模,从而提高文本分类的性能.具体而言,超图是普通图结构的扩展,由节点和超边组成,其中超边可以连接任意数量的节点.在 HyperGAT 模型中,文本超图的节点表示文本中的单词,节点属性为独热向量或预训练的词嵌入信息,超边分为序列超边和语义超边.对于序列超边,由于序列上下文描述了文本中单词之间的局部共现属性,且现有的研究已经证明其对于文本表示学习的有效性.为了利用该信息,HyperGAT 将文档中的每个句子视为超边,即序列超边,并与句子中的所有单词节点相连.此外,使用句子作为序列超边能够帮助模型同时捕捉文档的结构信息.此外,为了丰富每个单词的语义上下文,HyperGAT 构建语义超边以捕获单词之间与主题相关的高阶相关性.首先通过线性判别分析(Linear Discriminant Analysis,LDA)算法生成文档相关主题集合 T,其中每个主题 $t_i \in T$ 表示为在所有单词上的概率分布,即 $t_i=(\theta_1,\theta_2,\cdots,\theta_{|V|})$,$|V|$ 表示词汇表大小.随后将每个主题视为一个语义超边,与概率分布后降序得到的 top-K 个单词相连接.为了能够对所构建的文本超图处理,进行文本表征学习,HyperGAT 利用两个不同的聚合函数对节点图中节点表征进行学习,能够捕获文本超图上单词的异构高阶上下文信息,具体定义如下:

$$\boldsymbol{h}_i^l=\mathrm{AGGREGATION}_{\mathrm{edge}}^l(\boldsymbol{h}_i^{l-1},\{\boldsymbol{f}_j^l \mid \forall e_j \in \mathcal{E}_i\}) \tag{36}$$

$$\boldsymbol{f}_j^l=\mathrm{AGGREGATION}_{\mathrm{node}}^l(\{\boldsymbol{h}_k^{l-1} \mid \forall v_k \in e_j\}) \tag{37}$$

其中 $\mathcal{E}_i$ 表示与节点 v_i 相连的超边,$\boldsymbol{f}_j^l$ 表示在第 l 层超边 e_j 的表征信息;$\mathrm{AGGREGATION}_{\mathrm{edge}}$ 和 $\mathrm{AGGREGATION}_{\mathrm{node}}$ 分别表示聚合超边特征至节点以及聚合节点特征至超边的函数.

这两个函数主要通过对偶注意力机制来实现,包括节点级注意力机制和边级注意力机制.其中节点级注意力机制聚合得到超边 e_j 更新后的表征信息 $\boldsymbol{f}_j^l$,具体形式如下:

$$\boldsymbol{f}_j^l=\sigma\Big(\sum_{v_k \in e_j}\alpha_{jk}\boldsymbol{W}_1\boldsymbol{h}_k^{l-1}\Big) \tag{38}$$

$$\alpha_{jk}=\frac{\exp(\boldsymbol{a}_1^{\mathrm{T}}\boldsymbol{u}_k)}{\sum_{v_p \in e_j}\exp(\boldsymbol{a}_1^{\mathrm{T}}\boldsymbol{u}_p)} \tag{39}$$

$$\boldsymbol{u}_k=\mathrm{LeakyReLU}(\boldsymbol{W}_1\boldsymbol{h}_k^{l-1}) \tag{40}$$

其中 σ 是非线性激活函数,$\boldsymbol{W}_1$ 是网络参数矩阵,α_{jk} 表示节点 v_k 和超边 e_j 之间的注意力权重,$\boldsymbol{a}_1^{\mathrm{T}}$ 是随机生成的上下文向量.

边级注意力机制遵循相同的思想,但沿着相反的方向对边进行加权聚合以得到更新后的节点表征信息.这两种机制分别对传播过程中节点和边的不同影响力进行建模,从而能够学习到有效的图表征信息,用于最终的文本标签预测过程.实验结果表明,该模型中超图的构建以及相关网络的设计能够有效提高文本分类效果.其中,在新闻分类数据集 R8[155]上,该模型达到了 98.0%的准确率.

相较于其他模型,GNN 的信息层次传播聚合的思想在学习自然文本结构方面,尤其是远距离依赖关系捕获等问题上,具有得天独厚的优势.不过,GNN 相较于 RNN 等模型,由于全局信息引入而造成的噪声信息影响,难以精准捕捉文本的局部信息,尤其是小样本数据.故 GNN 通常着重考虑在需要捕捉文本结构性信息的场景使用.

4.4 基于迁移学习模型的分类方法

传统的深度学习方法通常基于此类前提假设:训练数据和测试数据源自同一领域,因此它们的输入特征空间和数据分布是一致的.然而,在实际应用场景如虚假新闻检测、新闻分类、主题标记等领域中,获取充足的训练数据往往困难重重,使得该假设难以成立.为应对该挑战,迁移学习(Transfer Learning,TL)[38]应运而生.迁移学习的核心思想是从相关任务中迁移知识,以此提高新任务的学习效果并减少对大量训练数据的依赖.迁移学习已在众多领域取得成功,如情感分类[156]、计算机视觉[157-159]等.特别在跨域、多层次文本分类任务上,迁移学习展现出了显著的成效.

2018 年,Howard 等人[99]提出了通用语言模型微调方法 ULMFiT,实现类似于 CV 的迁移学习方法,可应用于不同的 NLP 任务,而无需额外的数据补充.该方法在不同的文本分类任务上都取得了成功,且在仅有 100 个标注数据的小样本实验设置下

将误差减少了 23.7%. 2019 年，Banerjee 等人[100]针对多标签文本分类问题，将其分解为多个二分类问题，并提出了一种层次化迁移学习的策略 HTrans，其核心思想是层次结构中较低级别的二分类器是使用其父分类器的参数初始化的，并在子类别分类任务中进行了微调. 实验表明，相比于从零开始训练的二元分类器，该方法在 RCV1 数据集[140]上的 micro-$F1$ 和 macro-$F1$ 分数分别显著提高到了 0.76 和 0.48. 同年，Houlsby 等人[160]针对当前预训练语言模型在下游任务微调模型时参数学习效率低下的问题，提出一种参数迁移学习的方案. 其核心思想是针对每一个下游任务提供一个只有几个可训练参数的适配器模块，通过保持原始模型的参数不变，实现高度的参数共享. 因为适配器固定了原始预训练模型的参数，因此该模型只需优化 3.6%左右的参数，就可以达到和微调模型全部参数方法相似的性能. 2020 年，Raffel 等人[161]通过提出一个将所有基于文本的语言问题转换为文本到文本格式的统一框架，探索 NLP 迁移学习技术的前景. 2021 年，Cao 等人[101]针对跨域情感分类方法大多侧重挖掘源域和目标域间共有的知识，不能很好地利用目标域中未标记数据的问题，提出了一种用于细粒度跨域情感分类的深度转移学习机制. 该方法将不同域的文本数据映射到共享的特征空间中，同时利用 KL 散度设计域自适应模型，消除源域和目标域之间特征分布的差异，挖掘在跨域场景下可以自适应迁移的情感语义特征. 此外，考虑到现实社交场景下的文本多语种特性，2023 年 Unanue 等人[102]提出了 T3L 方法，将翻译和分类阶段分离，通过将目标语言翻译成高资源语言(如英语)，再使用高资源语言中训练的文本分类器进行分类. 在跨语言文本分类数据集上的实验表明，T3L 在绝大多数情况下优于跨语言基线模型，并在相关数据集 XNLI[162]上获得了 68.6%的预测准确率.

对比其它方法，迁移学习具有较弱的数据依赖性和标签依赖性，目前在文本分类领域是一个研究热点. 在探索文本分类领域下新的应用方案时，需要进一步解决复杂场景中的知识迁移问题，例如如何衡量跨域文本数据的可迁移性以及内容隐私问题. 另外，文本中的词语在不同的语境下可能具有不同的涵义，当从不太相关的源域中转移知识时，可能对目标域产生负面影响，如何应对此类负迁移现象亟需有效的解决方案.

4.5　基于预训练模型的分类方法

预训练模型(Pre-Trained Models，PTM)在自然语言处理中已得到广泛的应用[163]. 预训练技术旨在提前在大量数据上学习数据的通用特征，再将这些学习到的特征用于下游任务. 如在自然语言分类任务中，可以首先使用预训练模型在大量的文本数据上学习到语义特征、语法结构等通用信息，之后再针对下游分类任务对已学习到的模型使用少量训练数据，通过进一步训练，从而使得下游任务用户以较低成本获得较好的分类效果. 得益于预训练模型中学习了大量通用信息，使用其处理下游任务时往往具有模型收敛速度较快、数据依赖程度较低、性能较强且不易过拟合等优点.

4.5.1　常见预训练语言模型

预训练模型可以被分为有监督学习模型、无监督学习模型及自监督学习模型，其中有监督学习模型因为昂贵的数据标注成本在预训练模型中应用较少. 因此，本文着重对无监督学习模型与自监督学习模型进行介绍.

在预训练模型中，无监督学习的应用十分广泛. 这是由于预训练模型需要大量的数据来完成训练，使用无监督学习技术可以避免高昂的数据标注成本. 而自监督学习是有监督学习和无监督学习的一种综合，其核心思想是从大规模无标签数据中挖掘自身的监督信息. 如语言模型相关任务，通过基于输入的一部分数据来预测另一部分数据，可以学习得到对输入语义信息进行建模的模型. 该模型的学习过程和有监督学习几乎相同，不同之处在于其训练数据通常自动生成，模型直接从无标签数据中自行学习一个特征提取器，无需标注数据. 目前，自监督学习成功应用于大规模预训练语言模型，如 BERT、GPT、GPT2、GPT3 等.

Word2Vec[48]作为早期预训练语言模型的代表，便是使用无监督方式进行训练. 正如第 2.3 节所述，Word2Vec 主要通过两种思路实现：跳元模型(Skip-Gram)与连续词袋(Continuous Bag-Of-Words，CBOW)模型[27]. 两种模型均为词嵌入模型，即学习词汇向量表示，可以为文本分类任务提供更丰富、更有表现力的特征. 2020 年，Jang 等人[164]将 Word2Vec 给出的词嵌入作为 CNN 的输入来完成二者的结合，该模型在 IMDB Reviews 数据集[165]上达到了 91.4%的准确率. 但是，Word2Vec 所生成的词嵌入仅仅与词汇本身有关，而与词汇的具体上下文无关，这有可能对任务性能产生影响. 比如，在“苹果一斤五元”和“苹果手机有着强大的性能”这两个句子中，“苹果”这个词语有着不同的涵义. 然而，在完成预训练后，Word2Vec 只会给“苹果”这个词语唯一的词嵌入. 在文本分类等任务中，此类嵌入表示将极大影

响模型性能.

掩码语言模型. 针对以上问题,研究人员提出了基于掩码的预训练语言模型(Masked Language Model,MLM). 在该类模型的训练过程中,文本中的部分词语会被遮掩,模型则需要根据上下文对被遮掩的部分进行预测. 得益于这种训练方法,掩码语言模型不仅可以获取上下文感知能力,也可以避免人工标注的昂贵成本. BERT[30]是最流行的掩码语言模型之一. 与其他分类任务不同,文本分类通常需要考虑上下文信息,BERT 得益于 Transformer 架构的优越性能,通过双向编码可以对大量语境信息进行学习,得到了上下文感知的词嵌入表示,在进行文本分类任务时能够更好地理解文本的语义和语境. BERT 应用于文本分类时,需在输入序列的起始位置添加一个特殊标记[CLS],该标记用于产生序列分类的嵌入. 在进行文本分类时,只需将[CLS]最终产生的嵌入 $h_{[\mathrm{CLS}]}$ 输入一个简单的分类器(如 softmax),即可得到该序列的分类标签:

$$p(y_i|h)=\mathrm{softmax}(\boldsymbol{W}\boldsymbol{h}_{[\mathrm{CLS}]}) \tag{41}$$

其中 y_i 为标签,$\boldsymbol{W}$ 为权重矩阵,可以针对特定数据集进行训练学习.

基于 MLM 思想,2019 年 Sun 等人[109]针对文本分类任务提出了一种基于 BERT 的通用微调方案,其实验证明该方案在多个数据集上均达到了当时最好的指标. 2020 年,González-Carvajal 等人[166]在不同的文本分类场景中对比了 BERT 和传统模型的效果,其结果表明 BERT 在各个场景下,尤其是小规模数据集上,都优于传统方法. 例如,在 Chinese Hotel Reviews 这一较小数据集上,模型达到了93.8%的准确率.

自回归语言模型. 自回归语言模型(AutoRegressive Language Model,ARLM)是指在训练过程中,模型根据输入序列中先前单词预测下一个单词. 它是一种基于概率的模型,其主要思想是在给定词作为输入的情况下,根据当前输入生成下一个可能的词,并基于其出现的概率进行预测. 此外,ARLM 还可以反过来根据下文预测前面的单词. 具体地,给定一个长度为 n 的文本序列 $S=\{w_1,w_2,\cdots,w_n\}$,ARLM 的目标是学习一个条件概率分布 $\prod_{i=1}^{n}p(w_i|w_{<i})$ 或 $\prod_{i=n}^{1}p(w_i|w_{>i})$. 在文本分类任务中,文本表示是对文本的语义信息进行编码的关键. 自回归模型可以通过生成文本的过程来学习文本的表示,从而更好地捕捉文本的语义和特征[167],提高文本分类性能. 2018 年,OpenAI 团队提出的预训练模型 GPT[110]即为自回归语言模型,通过对 GPT 模型进行有监督微调后应用于文本分类任务,其在两种不同的文本分类任务上都超过了当前最佳结果. 在 GPT 的基础之上,OpenAI 研究人员逐步推出了 GPT-2[31]、GPT-3[32]、InstructGPT[111]以及 ChatGPT①. 2021 年,Edwards 等人[168]将 GPT-2 应用于小样本文本分类,区别于微调模型用于下游任务的方式,作者利用 GPT-2 基于原始样本生成额外的训练数据用于分类,从而提高小样本文本分类任务的模型性能. 2022 年,Yang 等人[112]针对当前预训练语言模型容易受到对抗性攻击的难题,提出一种鲁棒的前调(prefix-tuning)框架,其核心思想是在测试期间为原始前缀嵌入加入一个额外 batch 级前缀模块,以增强模型的鲁棒性. 在三个文本分类基准上进行的大量实验表明,该框架面对五种不同类型的文本攻击,相较于当前较好的基线模型显著提高了鲁棒性,同时在干净文本上保持了较高准确性. 2023 年,Sun 等人[169]利用 GPT-3 的文本生成能力提出了一种辅助大模型的渐进式文本推理策略 CARP,以解决文本分类中涉及的复杂语言现象. 该方法首先通过对话引导大模型寻找特定文本线索(例如,关键字、语气、语义关系、参考文献等),并在此基础上诱导模型推理以做出最终决策. 实验表明 CARP 在少样本场景下的表现让人印象深刻,仅使用每类 16 个样本 CARP 就在 SST-2[62]数据集上获得了 97.4%的准确率表现.

4.5.2 基于微调预训练模型的分类方法

目前,预训练语言模型在自然语言处理领域已经成为一个研究的热点. 这些模型通过在大规模的文本数据上预训练,学习语言的内在逻辑和通用表示,已在许多下游任务中展现出卓越的效果. 然而,这些通用表示并不能完全覆盖所有领域和任务的特定知识. 因此,研究者们开始探索如何将预训练语言模型扩展应用到更广泛的场景. 扩展预训练语言模型的一个重要方法是微调(Fine-tuning). 在微调过程中,预训练模型将与额外的网络进一步训练,以适应特定的任务或领域. 这种方法的特点在于能够保持预训练模型学习到的通用语言知识水平,同时融入特定任务或领域的专业知识. 相比于从头开始训练一个模型,微调预训练模型通常需要较少的数据和计算资源,这使得这种方法在面对现实场景时更加有效和可行. 在此,本文将概述几种基于微调预训练语言模型的文本分类方法.

① https://chat.openai.com

基于知识增强的预训练模型. 传统的预训练语言模型(如 BERT、GPT-2 等)使用海量的通用语料库进行训练,但是在特定领域的应用场景中,通用语料库往往难以满足需求. 针对以上挑战,研究人员尝试通过引入外部领域知识进行知识增强学习,常见的外部知识包括语料库、语义、常识、事实以及领域知识等. 在语料库和语义方面,有许多针对特定领域的词表和语料库,如医学领域的 BioBERT[170]、金融领域的 FinBERT[117]、法律领域的 LegalBERT[171]等. 模型通过利用特定领域的语料库进行训练,并在预测任务中使用领域特定的词表,以达到提高模型在特定领域表现的目的. 在领域常识和事实方面,研究人员尝试引入外部知识来增强模型的预训练效果. 如 2020 年 Ke 等人[116]提出了 SentiLARE 模型,通过利用 Sentiwordnet[172]为文本中的每个词提取上下文感知的情感极性,建立与句子情感之间的联系,达到将该信息融合到预训练任务中的目的,从而实现模型对语言知识的学习,提高模型情感分析任务表现.

在知识图谱领域方面,2020 年 Liu 等人[118]提出了 K-BERT 模型,通过将知识图谱中的实体和关系信息融入到模型中,以增强模型的知识表示能力. 另外,一些针对具体任务的模型也可以通过引入领域知识来增强模型的能力,如 2021 年 Sun 等人[119]提出了 ERNIE3.0 模型,通过结合百度百科和互动百科等大规模知识库来增强模型的常识和事实表示能力. 实验结果表明,ERNIE3.0 在 14 种自然语言理解任务上,包括情感分析、中文新闻分类和自然语言推理等,与之前的 10 个 SOTA 模型比较,取得了更优的性能表现. 其中在中文新闻数据集 THUCNews[173]上实现高达 98.7%的分类准确率.

基于对比学习的预训练模型. 对比学习(Contrastive Learning,CL)[174]是一种自监督学习方法,在不使用标签的情况下,利用数据之间的差异进行学习,其通过构建正样本(positive)和负样本(negative),然后度量正负样本的距离,让不同数据类型之间的表征区别开,同时拉近相同类型数据之间的表征. 2018 年 Google 的研究人员为解决 BERT 模型对句子级别的语境信息理解不足的问题,提出了 Next Sentence Prediction(NSP)任务[30]. 它是一种基于对比学习的解决方案,其目的是训练模型判断两个句子是否是连续的. 在构造数据集时,NSP 任务选取自然文本中相邻的两个句子 A 和 B 作为正样本,并采取随机采样的方式将句子 B 替换为语料库中任意一个其他句子作为负样本. 模型需要对两个句子进行编码表示,并判断两个句子之间的相似性,从而提高模型对于文本上下文关系的理解能力. 此外,基于对比学习的预训练模型可以有效提升模型在包括文本分类在内的各项任务的性能. 2019 年,Xiong 等人[121]提出了 WKLM 模型. 它利用维基百科,对原始输入中的实体进行替换,并通过对比学习训练模型对输入进行判别,让模型能够理解实体级别的文本特征. 实验表明,WKLM 模型对于许多下游任务,尤其是知识要求较高的任务,如语言问答、实体提取等有极大的增益. 在四个与实体相关的问题回答数据集上,WKLM 模型的表现始终优于 BERT,获得了 60.2%准确率表现. 与此类似,2021 年,Clark 等人[120]提出了 ELECTRA 模型. 与掩码语言模型通过特定嵌入遮盖输入并训练模型恢复不同,ELECTRA 模型从小型生成器中采样部分生成结果替换原始输入. 随后,ELECTRA 利用对比学习思想通过训练一个判别模型预测输入中的每个标记是否被生成器样本替换. 这样的训练方法可以让预训练模型能够学习到更高抽象层次的特征并且使用更少的参数. 实验表明,ELECTRA 模型在相同的模型大小、数据和计算量的情况下,相比于 BERT 模型平均准确率由 79.8%提升至 89.5%. 2021 年,Gao 等人[175]根据对比学习框架提出了可以用于无监督和有监督学习的 SimCSE 模型. 其中无监督 SimCSE 通过令同一个句子通过两次预训练模型,利用模型的两次 dropout 计算,得到两个不同的句向量作为正样本对;有监督 SimCSE 是利用自然语义推理(Natural Language Inference,NLI)数据集[176],把存在正确推理关系的句子对作为正样本对,矛盾的句子对作为负样本进行训练. 实验表明,与包括 BERT、Roberta 在内的现有预训练模型相比,SimCSE 在无监督和有监督场景中均有较大提升. 这得益于对比学习目标是将预训练嵌入的各向异性空间正则化,使其更加均匀,并且有监督信号可以进一步改善正例之间的对齐性,并产生更好的句子嵌入.

多模态预训练模型. 随着多媒体数据的不断增加,多模态(Multi-Modal,MM)预训练模型逐渐受到了研究者重视[177]. 多模态预训练模型是指能够同时处理来自不同模态数据(如文本、图像、音频等)的预训练模型. 相比于单模态预训练模型,多模态预训练模型可以利用不同模态之间的关联信息来提升模型的性能. 在文本分类领域,大多数工作通过结合文本和图像数据,试图使得模型学习到视觉和语义之间的联系,从而在多种任务上取得更好的表现. 例如 2019 年 Lu 等人[122]提出的 ViLBERT 模型,通过

一种双流机制，利用视觉和语言信息在 Transformer 层进行交互和融合，以学习视觉和语义之间的联系. 与 ViLBERT 不同，2019 年 Li 等人[123]提出的 VisualBERT 模型，利用单流的思想，通过将从目标建议中提取的图像特征转换为无序的输入 token，与文本一起作为模型的输入，由后续的多个 Transformer 层共同处理. 在这个过程中，词语和对象建议之间的丰富互动使该模型能够捕捉到文本和图像之间错综复杂的联系. 2022 年 Chen 等人[124]提出了 BEIT-3 模型. 该模型采用了 Multiway Transformer 骨干网络，通过将多种模态对应到不同的 Feed-Forward Network(FFN)参数与对 self-attention 进行参数共享，从而建立跨模态之间的语义关系，有效地建模了不同的视觉、视觉-语言任务. 与之前的多模态预训练模型采用多个预训练任务的方法不同，BEiT-3 模型通过将多模态数据都看作“语言”来实现预训练任务的统一. 这种统一的方法使得该模型能够广泛地支持各种下游任务，并有利于更大模型的训练. 此外，BEiT-3 模型的出现使得单流和双流模型有机地统一，为后续多模态预训练模型的发展提供了新的思路和方向. 实验结果表明，BEiT-3 在几乎所有与图像和语言相关的理解和生成任务中都达到了当前的最佳表现，包括但不限于语义分割、视觉问答、跨模态检索以及文本分类等任务.

基于微调预训练模型的分类方法通过额外的网络结构和训练，提取特定任务的数据特征，能够有效提高特定场景下的文本分类效果. 然而，由于其在文本分类领域的应用还处于探索阶段，仍存在一些缺点：(1)训练成本高，随着预训练模型参数量的逐渐增大，基于微调预训练模型的分类方法通常需要大量的计算资源来进行参数学习；(2)模型复杂化，为了适应不同领域的数据，这类方法需要设计特定的模型结构. 这不仅使得模型整体变得更加复杂，还需要一定的专业知识和经验来进行大量的参数调整和优化，以达到最佳性能；(3)可解释性较差，基于微调预训练模型的分类方法通常是黑盒模型，其决策过程难以解释，在某些场景下可能无法满足可解释性的需求.

4.5.3 提示学习

自 BERT 诞生以来，将预训练语言模型进行微调的学习方法已经成为了整个 NLP 领域的常规范式. 与需要引入额外的训练数据和使用特定目标函数的微调方法不同，提示学习(Prompt Learning)通过将下游任务转换为完形填空任务，随后由预训练语言模型填充空缺，并将填充符号映射至不同类别，从而最大限度地利用预训练语言模型，其主要思想与微调方法的对比如图 8 所示.

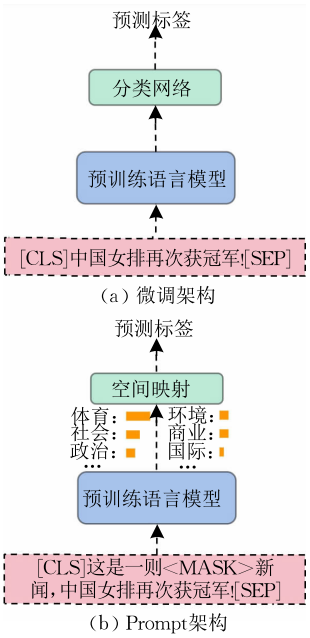

图 8 微调方法与 Prompt 方法的计算范式对比

给定文本一个长度为 n 的文本序列 $S=\{w_1, w_2, \cdots, w_n\}$，传统微调方法将文本 S 作为输入，通过预训练语言模型 PLM 提取文本特征矩阵 $\boldsymbol{M}$，进而使用分类网络 f 预测样本标签 y_i，其形式化如下：

$$\boldsymbol{M}=\mathrm{PLM}(S) \tag{42}$$

$$y_i=f(y_i|\boldsymbol{M}),\ y_i\in Y \tag{43}$$

而提示学习方法将一段额外文本 T 与 S 相连后，一同输入到预训练语言模型中进行预测，并通过映射空间 F 直接输出最终的预测标签：

$$y_i=F(y_i|\mathrm{PLM}(T,S)),\ y_i\in Y \tag{44}$$

以新闻分类任务为例[178]，给定新闻文本“中国女排再次获得冠军!”，提示学习的文本分类方法将句子转换为“**这是一则〈*MASK*〉新闻**，中国女排再次获得冠军!”，其中加粗的额外文本被称作提示(Prompt). 随后使用预训练模型在〈***MASK***〉处预测单词，如“体育”、“社会”、“国际”等，进而将预测单词映射至标签空间，最终完成新闻分类. 通过选择适当的提示，该方法可以将不同的下游任务快速转换，使预训练语言模型本身推理预测所需的输出，有时甚至不需要任何额外的特定训练，就可以达到一定的效果[179]. 本小节将对提示学习方法进行介绍，并对基于提示学习的文本分类方法[180-182]进行总结.

近几年提示学习方法在数据约束场景下的文本分类任务中展现出了强大的竞争力. 2021 年，Schick 等人[113]针对少样本文本分类场景，提出了一种通过词掩码将文本分类任务转换为完形填空任务的半

监督训练方法 PET. 实验表明仅使用 RoBERTa-base 模型，该方法就在多个文本分类任务下取得了当时最好的分类效果，并远远优于大多数监督和半监督方法，在仅有 10 个训练样本的实验设置下，PET 在 Yelp[183] 数据集上获得了 57.6% 的准确率. 考虑到 PET 方法人工设计提示的复杂性以及预训练模型对于不同提示的敏感性，2021 年 Liu 等人[114] 提出了名为 P-tuning 的方法，成功地实现了提示的自动构建. 该方法使用可学习的 token 加入预训练语言模型的输入之中，取代了 PET 方法里人为设计的固定提示，并允许其与预训练模型共同微调. 实验结果表明 P-tuning 方法大大降低了所需的人工设计成本，同时在 SuperGLUE[184] 数据约束设置下的实验结果显示，P-tuning 相比于 PET 方法提高了近 6.0% 的性能，获得了 79.1% 的平均准确率. 针对不同任务和不同预训练语言模型的设置需求不同，2022 年 Deng 等人[115] 提出了一种基于强化学习的高效离散提示搜索框架 RLPrompt. 通过使用 SQL[185] 算法联合训练的策略网络，RLPrompt 可以灵活适用于不同类型的预训练语言模型，并自动为其生成提示模板. 实验表明该方法优于传统微调、PET 以及 P-tuning 在内的多种文本分类方法，并且对不同的语言模型选择具有鲁棒性. 类似地，考虑到样本数据差异性，2023 年 Li 等人[186] 利用强化学习提出了一种基于策略梯度[187] 的样本级提示优化方法 DP_2O. 该方法利用 GPT-4[188] 生成提示，并通过策略网络为每个样本匹配合适的提示，在保证提示可读性的同时，提高了预训练模型在多个文本分类任务上的表现. 实验结果显示，在 SST-2[62] 数据集上，DP_2O 获得了 93% 的平均准确率，同时其训练时间仅为 RLPrompt 的 10.9%.

针对跨域文本分类问题，2021 年 Ben-David 等人[189] 利用元学习思想提出了一种跨域自适应生成提示的方法 PADA. 面对未知的目标域样本，PADA 根据训练时的源域数据，提取特征词汇，组合生成样本唯一的提示以增强预训练语言模型 T5(Text to Text Transfer Transformer)[161] 的跨域泛化能力. 实验结果表明 PADA 方法性能在多数设置下优于所有基线模型，并在 PHEME[190] 和 MNLI[191] 数据集上分别比使用传统微调方法训练的最佳 T5 模型表现出 3.5% 和 1.3% 的平均性能增益，获得了 69.3% 和 79.6% 的平均准确率. 另一方面，为应对标签空间与词汇空间映射损失的难题，2021 年 Hu 等人[192] 将外部知识融入到语言模型中，利用外部知识库扩展模型的词汇空间并进行降噪处理，进而使用扩展后的词汇空间进行预测，取得了更好的文本分类效果. 在零样本学习的场景中，包含 AG News[60] 等不同数据集设置下，该方法相比于最优基线 P-tuning 方法的错误率平均降低了 9.7%，获得了平均 82.6% 的准确率表现.

综上所述，与以往的基于预训练语言模型的分类方法相比，提示学习方法在低资源和少样本的文本分类场景拥有着得天独厚的优势. 然而对于以 PET、RLPrompt、DP_2O 为代表的离散提示优化方法，由于预训练语言模型对不同提示的高度敏感性，如何避免人工设计，快速准确地检索出与下游任务匹配的最优提示依然是一项巨大的挑战. 针对该挑战，目前的很多工作，例如 P-tuning，都尝试引入额外的模型结构自适应地生成提示，但如何平衡额外产生的计算资源消耗依旧是一个值得探讨的问题.

总体而言，深度学习作为目前最主流的人工智能实现方法，通过自适应地提取出分类样本的内在特征，克服了人工特征设计造成的不完备性缺陷，因此基于深度学习的分类模型在众多文本分类任务中大放异彩. 随着深度学习的不断发展，多样化的文本分类方法纷纷涌现，以应对该领域面临的各项挑战. 基于基础网络的分类方法摆脱了传统机器学习方法依赖的手动特征工程，从不同角度对文本数据进行建模，提取文本的多维特征，而基于胶囊网络和图神经网络的分类方法则分别针对文本数据长度的不一致性和文本分类的多层次、多粒度需求提供了解决方案. 此外，基于迁移学习和预训练模型的文本分类方法能够有效处理跨域、低资源和少样本等场景下的文本分类问题. 研究表明，在满足特定条件的应用场景下，大多数深度学习方法已经超越了现有传统机器学习方法的文本分类性能. 但是，为了达到较好的文本分类精度，一方面深度学习需要大量的训练数据支撑，另一方面由于算法模型的逐渐复杂化，其训练的运行时间与计算资源代价急剧攀升. 在这种情况下，深度学习方法所需的人工编程技巧和服务器硬件支持越来越高，这大大阻碍了深度学习方法的落地应用与推广. 因此，基于深度学习的文本分类方法在未来将朝着少样本、领域自适应、轻量级预训练模型等方向继续发展.

5　数据集

本节在已有工作[33-36] 基础上，对广泛应用于文本分类研究的数据集进行了系统性调研，根据其应用领域的不同对数据集进行了划分，并在表 3 给出了概述.

表 3 不同文本分类应用领域常用数据集统计汇总

应用领域	数据集名称	数据集规模	类别数量	平均文本长度	任务类型	语言
情感分析	Movie Review(MR)[193]	10662	2	20	二分类	英文
	Stanford Sentiment Treebank(SST-2)[62]	9613	2	19	二分类	英文
	Multi-Perspective Question Answering(MPQA)[194]	10606	2	3	二分类	英文
	IMDB Reviews[195]	50000	2	294	二分类	英文
	Yelp Reviews(Yelp-2)[183]	598000	2	153	二分类	英文
	Amazon Review(AM-2)[196]	4000000	2	91	二分类	英文
	Ren-CECps[197]	1487	8	591	多标签	中文
	Dianping[197]	2500000	2	142	二分类	中文
	BDCI2019[198]	14716	2	1204	二分类	中文
	NLPCC2014-SC[199]	54000	2	38	多标签	中文
	SE-ABSA16-PHNS[200]	1862	2	712	二分类	中文
	SE-ABSA16-CAME[200]	1740	2	704	二分类	中文
	Weibo_senti_100k[201]	100000	2	43	二分类	中文
新闻分类	20 Newsgroups(20NG)[79]	18846	20	221	多分类	英文
	AG News(AG)[60]	31900	4	235	多分类	英文
	Reuters Corpus(R8)[155]	7674	8	66	多分类	英文
	RCV1-2[140]	804414	2	124	多标签	英文
	THUCNews[202]	740000	14	964	多分类	中文
	Sogou News(sogou)[203]	2909551	5	—	多分类	中文
	ChinaNews[204]	1512000	7	119	多分类	中文
	TNEWS[205]	382688	15	44	多分类	中文
	Chinese News Same Event dataset[206]	29063	2	734	二分类	中文
	IFLYTEK[207]	17332	119	—	多分类	中文
虚假新闻检测	Chinese Rumor Dataset[208]	31669	2	110	二分类	中文
	LIAR[209]	12836	6	106	多分类	英文
	BuzzFeedNews[210]	1627	2	—	二分类	英文
	CREDBANK[211]	60000000	5	—	多分类	英文
	FacebookHoax[212]	15500	2	148	二分类	英文
	FakeNewsNet[213]	3000000	2	—	二分类	英文
	Twitter and Weibo DataSet[214]	5000	2	94	二分类	多语言
	Buzzfeed Election Dataset and Political News Dataset[215]	75	3	2353	多分类	英文
抽取式问答系统	Stanford Question Answering Dataset(SQuAD)[216]	107785	—	5000	二分类	英文
	TREC-QA(TREC-6)[136]	5952	6	1162	二分类	英文
	WikiQA[217]	3047	—	873	二分类	英文
	Quora[218]	404290	2	11	二分类	英文
	CR[219]	3775	2	19	二分类	英文
	cMedQA[220]	54000	—	182	二分类	中文
	webMedQA[221]	63284	—	137	二分类	中文
	XQA[222]	90610	—	361	二分类	多语言
自然语言推理	Stanford Natural Language Inference(SNLI)[223]	570152	3	11	多分类	英文
	Multi-Genre Natural Language Inference(MNLI)[224]	433000	3	16	多分类	英文
	Sentences Involving Compositional Knowledge(SICK)[225]	10000	3	10	多分类	英文
	Microsoft Research Paraphrase(MSRP)[226]	5800	2	22	多分类	英文
	SciTail[227]	27026	2	21	多分类	英文
	OCNLI[228]	56000	3	15	多分类	中文
	XNLI[162]	112500	3	—	多分类	多语言
主题标记	DBPedia[145]	630000	14	55	多标签	多语言
	Obsumed[229]	7400	23	136	多标签	英文
	Yahoo answers(YahooA)[61]	1460000	10	112	多分类	英文
	EUR-Lex[230]	19314	3956	1239	多标签	多语言
	Web Of Science(WOS-46985)[231]	46985	134	—	多标签	英文
	Fudan[232]	18655	20	2981	多分类	中文
	AAPD[233]	55840	54	163	多标签	中文
文本自动编码分类	MIMIC-III[234]	8922	65132	1500	多标签	英文
	CCHMC[235]	1954	45	21	多分类	英文
	Xiangya[236]	7732	1177	610	多标签	中文

5.1 情感分析(Sentiment Analysis,SA)

情感分析是指对带有情感色彩的主观性文本进行分析、处理、归纳和推理的过程,通常被应用于了解人们是否支持文本中的特定观点.情感分析相较于其他文本分类任务,要求模型对语境有着理解能力,可以分辨出文本数据中细微的情感差异[237-240].利用情感分析能力,可以针对带有主观描述的自然语言文本,自动判断该文本的情感两极性倾向并给出相应的结果,即文字中表述的观点是积极的或者消极的.此外,利用"超出两极性"的情感分析能够对更复杂的情绪进行判断,例如"生气"、"悲伤"、"快乐"等等.常见的SA数据集包括 Movie Review(MR)[193,241]、Stanford Sentiment Treebank (SST)[62,83]、Multi-Perspective Question Answering (MPQA)[165,194]、IMDB[165]、Yelp[183]、Amazon Reviews (AM)[138]、Subj[242]、CR[243]、SS-Twitter[244]、SS-YouTube[244]、SE1604[245]和 Dianping[246]等.下面本文将对部分数据集进行详细介绍.

Movie Review[193,241].这是2002年公开的电影评论数据集,每条数据为依据观念状态(主观/客观)或情感极性标记的评论句子.数据集共包含5331条积极情感和5331条消极情感的句子/片段.在实验设置中,通常基于十折交叉验证对数据集进行随机划分,从而评估模型的预测效果.

Ren-CECps[247].这是一个由中文博客构成的多标签情感分类数据集,其中的数据被分为愤怒、焦虑、期待、恨、爱、快乐、悲伤、惊吓等8种不同的情感类别,以及正面、负面、中性3种极性.该数据集包含了1487篇文档、11 255个段落、35 096条句子以及878 164个词语.

Stanford Sentiment Treebankcite[62,83].这是斯坦福大学公开的情感分析数据集,主要针对电影评论来做情感分类,因此该数据集可以用于单个句子的文本分类任务.其包含两类数据:(1)SST-2是二分类数据,包含9613个被标记文本,分为6920个训练文本、872个验证文本和1821个测试文本;(2)SST-5是五分类数据,情感极性区分得更为细致,包含8544个训练文本和2210个测试文本.

Multi-Perspective Question Answering[165,194].这是由MITRE公司公开的多视角问答数据集,其中包含对观点极性检测的相关子任务.数据集共包括10 606条从不同来源的新闻中所提取的句子.同时,这是一个不均衡数据集,包含3311条正样本和7293条负样本.

IMBD Reviews[165].该数据集包含50 000条电影评论,可以用于自然语言处理或文本分析.其中每条评论被标注为正面/负面两类情感极性,每类评论的数量相同,使用过程中可以被均分为训练组和测试组,每组包含25 000条评论数据.

Yelp Reviews[183].该数据集是Yelp业务、评论和用户数据的子集,最初是为Yelp Dataset Challenge而设计.其包含两类数据:(1)Yelp-2用于二元情感分析任务,包含560 000条训练文本和38 000条测试文本;(2)Yelp-5用于进行更细粒度的情感分析,所有类别中共有650 000条训练文本和50 000条测试文本.

Amazon Reviews[138].此数据集来自亚马逊网站的产品评论和元数据.其包含两类数据:(1)Amazon-2有2个类别标签,包含3 600 000条训练文本和400 000条测试文本;(2)Amazon-5有5个类别标签,包含300万条训练文本和65万条测试文本.

BDCI2019[198].该数据集用于2019年CCF大数据与计算智能大赛中互联网情感分析赛题,包括新闻网、微信、博客、贴吧等网站所爬取的金融相关文本数据,共包含7349条训练文本和7367条测试文本.

NLPCC2014-SC[199].该数据源自新浪微博,用于NLPCC 2014 Shared Task中情感分析任务,对于输入的整条微博,任务要求判断出该微博是否包含情绪.对包含情绪的微博,任务要求判别其情绪类别,包括愤怒、厌恶、恐惧、高兴、喜好、悲伤、惊讶等.由于一条微博中可能包含多个不同个体与情绪,数据集中每条文本都被标注了两种主导情绪.该数据集共包含14 000条训练数据和40 000条测试数据.

Weibo_senti_100k[201].该数据集包含10万条新浪微博文本数据,平均长度为42.9个字,包括积极情绪和消极情绪两类情感标签,其中每个类别大约有50 000条数据.

SE-ABSA16-PHNS/CAME[200].该数据集由哈尔滨工业大学的研究人员整理,用于SemEval-2016中的情感分析任务.数据集由客户评论和相应人工标注组成,标注包含了评论的主要方面(aspect)和相应情感正负.后缀PHNS指手机相关评论,共有1862条句子级标注;CAME指数码相机相关评论,共有1740条句子级标注.

5.2 新闻分类(News Categorization,NC)

新闻是最重要的信息获取渠道之一,对人们生活有着至关重要的影响.新闻分类是文本挖掘领域

较为常见的应用场景，有利于实现人们对重要信息的筛选. 相较于其他文本分类任务，新闻分类要求模型可以处理广泛的文本主题和多模态信息，且常常涉及到多标签、多粒度的分类任务[63,248-249]. 新闻分类的应用主要包括识别新闻主题、新闻极性分析和根据用户兴趣推荐相关新闻等. 常见的新闻分类数据集包括 20 Newsgroups(20NG)[79]、AG News(AG)[60]、Reuters Corpus[155]、THUCNews[173]、Sogou News(Sogou)[230] 和 ChinaNews[246] 等. 下面本文将对部分数据集进行详细介绍.

20 Newsgroups[79]. 此数据集是用于文本分类、文本挖掘和信息检索研究的国际标准数据集之一. 它是由 18 846 个新闻组文档构成的数据集合，在 20 个不同类别的新闻组中几乎均匀划分.

AG News[60]. 此数据集是 AG 新闻文章语料库的子数据集，由 AG 语料库的四大类(“世界”、“体育”、“商业”、“科技”)文章的标题和描述字段组合而成. AG 中每个类别包含 30 000 条训练文本和 1900 条测试文本.

TNEWS[205]. 该数据集源自今日头条 2018 年 5 月之前发布的中文新闻，共包含 382 688 条新闻，涵盖了 15 个类别的新闻，包括旅游、教育、金融、军事等.

IFLYTEK[207]. 该数据集是由科大讯飞于 2019 年创建的竞赛数据集，用于应用类别预测. 该数据集包含 17 332 条应用程序的描述文本数据，涵盖 119 种和日常生活相关的应用程序类别，例如食物、租车、教育等.

RCV1-2[140]. 该数据集是一个多标签英文文本分类数据集，由路透社所提供的新闻数据集 RCV1 修正而来. 它包含了共 804 414 条数据以及 103 种主题.

Chinese News Same Event Dataset[206]. 该数据集由专业编辑进行标注，由中国主要互联网新闻提供商发表的长篇中文新闻构成，涵盖了开放领域的各种主题. 数据集共包含 29 063 条新闻文章对，标签用以标注一对新闻文章是否报道了同一件突发新闻事件.

Reuters Corpus[155]. 此数据集是一个英文新闻语料数据，这些新闻出现在 1987 年的路透社通讯上，由路透社公司的人员进行手工分类标注. 该数据集是从源数据中提取高频类别下的新闻所构成，主要包括：(1) R8 有 8 个类别标签，包含 5485 条训练文本和 2189 条测试文本；(2) R52 有 52 个类别标签，包含 6532 条训练文本和 2568 条测试文本.

THUCNews[173]. 此数据集是根据新浪新闻 RSS 订阅频道 2005～2011 年间的历史数据筛选过滤生成，包含 74 万篇新闻文档. 并在原始新浪新闻分类体系的基础上，重新整合划分出 14 个候选新闻类别标签，分别是财经、彩票、房产、股票、家居、教育、科技、社会、时尚、时政、体育、星座、游戏和娱乐.

Sogou News[250]. 此数据集由 SogouCA 和 SogouCS 新闻语料库构成，具有 5 个类别共计 2 909 551 篇文章，每个类别均包含 90 000 个训练样本和 12 000 个测试样本，每个样本的类别标签都是 URL 中的域名，并且这些样本均已转换为拼音.

5.3 虚假新闻检测(Fake News Detection，FND)

随着社交媒体的革命性发展，人们逐渐习惯在 Twitter、Facebook、微博、微信等不同的社交媒体平台阅读、发布和传播新闻. 但是，在社交媒体带来生活便利的同时，也使得未经证实的虚假新闻得以大规模的出现并快速传播，这严重危害着公众对政府和新闻业的信任，深深影响人们的正常生活. 因此，在社交媒体平台中对未经证实的新闻加以验证、检测并清理虚假新闻变得至关重要. 虚假新闻检测在新闻分类的基础上，往往还要求分类模型可以协同处理各种社交信息(如新闻本身、评论、转发等)[251-254]. 常见的虚假新闻检测数据集包括 Chinese Rumor Dataset[208,255]、LIAR[209]、BuzzFeedNews[210]、CREDBANK[211]、FacebookHoax[212]、FakeNewsNet[251,256-257]、Twitter and Weibo DataS[258]、Buzzfeed Election Dataset and Political News Dataset[259] 等. 下面本文将对部分数据集进行详细介绍.

Chinese Rumor Dataset[208,255]. 该数据集是由清华大学自然语言处理实验室于 2018 年发布，其数据来源于新浪微博不实信息举报平台从 2009 年 9 月 4 日至 2017 年 6 月 12 日提供的 31 669 条中文虚假新闻数据. 该数据集分为两个部分，分别包含了虚假新闻原微博与其相关的转发、评论等信息.

BuzzFeedNews[210]. 该数据集是由 9 家新闻机构提供的从 2016 年 9 月 19 日至 23 日，即美国大选前一周，在 Facebook 发布的完整新闻样本. 它包含了 1627 篇文章，其中 826 篇来自主流媒体，如 CNN、ABC News 等，256 篇来自左翼媒体，545 篇来自右翼媒体. 数据集中每篇新闻样本都由至少 5 名 BuzzFeed 工作人员进行了事实核查，以确保新闻标签的准确性.

LIAR[209]. 该数据集来自于实核查网站 PolitiFact 收录的 12 800 名用户的短文本陈述，包含了医

疗、政治、国际、环境等主题. 该数据集数据被分为谎话连篇、完全虚假、半真半假、勉强真实、大部分真实、完全真实等 6 个大类.

FakeNewsNet[251,256-257]. 该数据集由从事实核查网站 PolitiFact 和 GossipCop 收集的虚假和真实新闻组成. 其数据不但收录了新闻文章的原版内容，还包含了与新闻文章相关的用户反馈，例如用户回复、转发和收藏等. 数据集总数包含了近 200 万条与虚假和真实新闻相关的推文数据.

CREDBANK[211]. 该数据集包含了从 2015 年 10 月开始 96 天内大约 6000 万条新闻推文，其被划分到 1049 个真实世界中的事件. 为保证事件的可靠性，每个事件都由来自 Amazon Mechanical Turk 的 30 名工作人员进行评估并进行标注.

5.4　抽取式问答系统(Extractive Question Answer, EQA)

基于阅读理解的问答系统在 NLP 领域近年来备受关注，该过程又可以分为抽取式问答和生成式问答. 其中，抽取式问答的原理是给定一个问题和一段文本，机器自动在阅读理解的基础上，从这段文本中找出能回答该问题的文本片段，实质为对所提供的多个候选答案依次进行分类判断其是否属于正确答案的过程. 所以抽取式问答系统可以应用文本分类模型来识别出正确答案，并将其它相关答案设置为候选答案. 常见的数据集包括 Stanford Question Answering Dataset (SQuAD)[216,260]、TREC-QA[136]、WikiQA[217,261]、Subj[242]、CR[243]、Quora[218]、cMedQA[262]、webMedQA[263]、XQA[222,264]等. 下面本文将对部分数据集进行详细介绍.

Stanford Question Answering Dataset[216,260]. 此数据集是斯坦福大学于 2016 年推出的一个大规模英文阅读理解语料数据集，由一系列维基百科文章的提问和对应的答案构成，其中每个问题的答案都可以从相关文章中的文本片段或区间找到. 该数据集共包含 107 785 个问题以及配套的 536 篇文章.

TREC-QA[136]. 此数据集是问答任务研究中最常用的英文数据集之一，包含 TREC-6 和 TREC-50 两个版本. 其中，TREC-6 包含 6 个类别的问题，而 TREC-50 包含 50 个类别. 此外，这两个版本内的数据分布一致，均包含 5452 个训练问题和 500 个测试问题.

WikiQA[217,261]. 此数据集是微软研究院于 2016 年推出的一组公开的问题和句子对构成的英文语料库，用于开放领域问答的研究. 数据集中每个问题都链接到维基百科的一个可能含有答案的页面. 该数据集共包含 3047 个问题和 29 258 个句子，其中 1473 个句子被标记为相应问题的答案句. 此外，该数据集也包括了一些没有正确答案的问题.

cMedQA[262]. 此数据集是国防科技大学团队于 2017 年推出的一组公开的中文在线医学问答匹配语料库，包含 cMedQA 和 cMedQA2 两个版本. 其中，cMedQA 包含 54 000 个问题及对应的约 101 000 个回答；cMedQA2 是 cMedQA 的扩展版，包含约 10 万个医学相关问题及对应的约 20 万个回答.

XQA[222,264]. 此数据集是清华大学团队于 2019 年针对开放式问答构建的一个跨语言数据集，包含英语的训练集以及英语、法语、中文等九种语言的验证集和测试集. 其中，训练集包含 56 279 对英语问答对以及相关文档，验证集和测试集分别包含 17 358 和 16 973 对问答对. 所有问题都来自以相应语言为母语的人，并潜在地反映了不同语言的文化差异.

5.5　自然语言推理(Natural Language Inference, NLI)

自然语言推理，又称文本蕴含识别(Recognizing Textual Entailment, RTE)，是自然语言处理中常用的一类文本分类任务，指的是判断一段文本的意思是否蕴含于另一段文本. 其中，复述检测(paraphrase identification)是自然语言推理的一种广义形式，目的是判断两个句子是否有相同含义. 与其他文本分类任务只需要将文本映射到一个或几个固定的类别不同，NLI 任务的挑战在于理解和推断文本之间的复杂关系，这要求分类模型具有逻辑、常识推理和语言理解能力[265-267]. 常用的英文自然语言推理数据集包括 Sentences Involving Compositional Knowledge (SICK)[225,268]、Stanford Natural Language Inference (SNLI)[269]、Multi-Genre Natural Language Inference (MNLI)[191,224]、Microsoft Research Paraphrase (MSRP)[226,270]、Recognising Textual Entailment (RTE)[271]和 SciTail[272]等. 常用的中文自然语言推理数据集包括 Original Chinese Natural Language Inference (OCNLI)[228,273]. 此外，还有跨语言的自然语言推理数据集 Cross-Lingual Natural Language Inference (XNLI)[196]. 下面本文将对部分数据集进行详细介绍.

Sentences Involving Compositional Knowledge[225,268]. 该数据集是 Marelli 等人于 2014 年提出，由两个已经存在的数据集 8K ImageFlickr 和 SemEval 2012 STS MSR-Video Description 衍生而来的集合，包含 10 000 条利用蕴含类别标签标注的句子对，其中蕴

含、中立、矛盾三种类别标签分别包含 2821、5595 和 1424 个.

Stanford Natural Language Inference[269]. 此数据集是斯坦福大学于 2015 年推出的一个大规模人工标注的英文句子对的集合,被广泛应用于自然语言推理任务.它包含 570152 条人工标注的句子对,其中每对句子之间的关系都用中立(neutral)、蕴含(entailment)和矛盾(contradiction)三个标签中的一个进行标记.

Multi-Genre Natural Language Inference[191,224]. 此数据集是纽约大学于 2017 年推出的一个多类型自然语言推理数据库,是通过众包方式对句子对进行文本蕴含标注的集合.数据集分为 matched 和 mismatched 两个版本,其中 matched 指的是训练集和测试集的数据来源一致,mismached 指的是训练集和测试集的数据来源不一致.该数据集共包含 433000 个句子对,给定了前提语句和假设语句,共有 3 种类别标签,分别为蕴含假设、与假设矛盾以及中立.该数据集和 SNLI 比较类似,但是前提语句来自真实场景的数据,是从数十种不同来源收集的,包括转录的语音、小说和政府报告等.

Original Chinese Natural Language Inference[228,273]. 此数据集是 Hu 等人于 2020 年推出的一个大型原生中文自然语言推理数据集,不依赖于任何机器翻译或非专家注释,包含 56000 个带注释的句子对.该数据集给定了前提语句和假设语句,其中前提是从政府文件、新闻、文学、电视谈话节目和电话会话五个文本类型中收集,共有 3 种类别标签,分别为蕴含、中立和矛盾.

Cross-Lingual Natural Language Inference[176]. 此数据集是 Facebook AI 研究院和纽约大学研究团队于 2018 年合作推出的一个跨语言自然语言推理数据库,支持英语、法语、汉语等 15 种语言,包含 10 个领域.数据集从 MNLI 使用的 10 种来源中分别收集并验证了 750 个新实例,再将这 7500 个实例分别翻译成其他 14 种语言,产生了 112500 对句子对.其中,每条数据由前提和假设两个句子组成,有 3 种类别标签,分别为蕴含、中立以及矛盾.

5.6 主题标记(Topic Labeling,TL)

统计主题由于其广泛的应用,近年来在机器学习和文本挖掘中被广泛关注.在主题标记领域,其主要的挑战是发现文本数据中潜在的主题或话题,准确地为每个文档指定一个或多个主题,以便用户能够解释发现的主题并进行分析.常用的主题标记数据集包括 Ohsumed[229]、DBPedia[145,174]、EUR-Lex[230]、Yahoo Answers (YahooA)[61]、Web Of Science (WOS)[275]、Amazon670K[276]、Bing[277] 和 Fudan[232] 等.第 5.6 节将对部分数据集进行详细介绍.

Ohsumed[229]. 此数据集来源于在线医学信息数据库 MEDLINE 的 348566 篇参考文献,由 1987 年至 1991 年五年间的 270 种医学期刊的标题和/或摘要组成,共包含 37400 个文档.可用字段包括标题、摘要、网格索引术语、作者、来源和出版物类型.

DBPedia[145,274]. DBpedia 从 111 种语言的维基百科版本中提取结构化数据,构建了一个大规模的多语言知识库,涵盖了广泛的主题.其中最流行的版本有 14 个类别标签,分别包括 560000 条训练数据和 70000 条测试数据.

EUR-Lex[230]. 此数据集是关于欧盟法律的文件集,涉及欧洲法律的不同方面,根据正交分类方案进行索引.其中包含许多不同类型的文件,例如条约、立法、判例法和立法提案等.该数据集的最常用版本收集了 19314 个文档和 3956 个类别标签.

YahooA[61]. 此数据集是使用原始 Yahoo Answers 语料库中抽取 10 个最大的主要类别数据所构建的,应用于主题标记任务.其中每个类别包含 140000 条训练样本和 6000 条测试样本,每条文本都包含问题标题、问题上下文和最佳答案三个元素.

WOS[275]. 此数据集收集了 Web of Science 上收录的文章.WOS 作为独立于发行商的全球引文数据库已发布三个版本:WOS-46985、WOS-11967 以及 WOS-5736.其中 WOS-46985 是完整的数据集,共包含 46985 个文档,WOS-11967 和 WOS-5736-是该完整数据集的子集.

BDCI-2018[278]. BDCI-2018 是一个汽车行业用户观点主题及情感识别任务数据集,收录了汽车论坛中的评论,其中训练数据集 12572 条,测试数据集 2753 条.训练集中的数据根据以下 10 个主题进行划分:动力、价格、内饰、配置、安全性、外观、操控、油耗、空间、舒适性.该数据集情感分为 3 类,包括中立、正向和负向.

5.7 文本自动编码分类(Automated Coding,AC)

文本自动编码分类指将原始文本中对特定客体的不规范表述与该客体的标准编码相关联.该任务广泛应用于医疗领域,实现对病例中包含的疾病、药物等重要信息的自动编码.与其他文本分类任务相比,文本自动编码分类因与专业领域的相关性,往往需要专家知识对模型设计进行指导.目前已有很多

工作对该问题进行研究[279-282]. 本文对目前较为常见的文本自动编码分类数据集进行了详细介绍.

MIMIC-III[234]. 这是一个重症医学数据集，其由 2001 年 6 月至 2012 年 10 月之间 61 532 份病例构成，包含了对于疾病、医疗过程等内容的 ICD-9 编码描述以及相应的 ICD-9 词典表，可用于多标签的医学文本自动编码. MIMIC 数据集目前仍在持续更新中.

CCHMC[235]. 该数据集收集了来自辛辛那提儿童医院医学中心放射科的医疗数据，在医学自动编码的研究中被广泛使用. 该数据集共有 1954 份放射结果报告，对应着 45 种医学编码. 该数据集使用的编码版本为 ICD-9-CM，其中每份文档只对应于一种医学编码.

Xiangya[236]. 该数据集收集了中南大学三家附属医院的电子健康病例，共计 7732 份文档. 该数据集包含了 1177 种医学编码，每份文档平均对应 3.6 个医学编码，其使用的编码类型为 ICD-10.

6　评价方法

在文本分类任务中，最初常用的评价方法是准确率和 $F1$ 分数. 近些年来，由于文本分类任务复杂度的提高以及分类任务应用场景的多样化，对应的评价方法随之不断改进，各种各样的评价方法应运而生. 例如，评估指标 $Micro\text{-}F1$ 和 $P@K$ 常用于对多标签文本分类算法的性能进行评估，EM 则被用于评估问答系统的性能. 因此，为了进一步对不同分类模型的性能进行评估对比，需要了解不同评价指标的基本机制以及这些指标中所蕴含的物理意义和结果信息. 本节将对常用的评价指标和性能度量方法进行讨论. 为方便阅读，在表 4 中，本文给出评价指标中常用的符号说明.

表 4　评价指标中使用的符号定义

符号	表示含义
TP	真正类，实际为正类
FP	假正类，实际为负类
TN	真负类，实际为负类
FN	假负类，实际为正类
TP_t	第 t 个类别的真正类
FP_t	第 t 个类别的假正类
TN_t	第 t 个类别的真负类
FN_n	第 n 个类别的假负类
Y	类别集合
N	样本数量
Q	查询数量

6.1　单标签评价指标

在最简单的情况下，单个样本只属于一个标签类别. 这种情况下，文本分类可以视为一个多分类（含二分类）任务[283]. 下文对常见的单标签评价指标进行介绍.

准确率（*Accuracy*）和错误率（*Error Rate*）. 作为分类任务最基本的指标，准确率和错误率常被用来评价单标签文本分类模型的性能[284]. 它直接反映了分类准确和错误的样本各自所占的比例，具体如下：

$$Accuracy=\frac{TP+TN}{N} \tag{45}$$

$$Error\ Rate=1-Accuracy=\frac{FP+FN}{N} \tag{46}$$

其中 N 表示样本总数量.

该评价指标计算简单，易于理解，应用十分广泛，但是当数据不均衡时，其结果无法很好的衡量分类器的好坏.

灵敏度（*Sensitivity*）和特异度（*Specificity*）. 灵敏度对应于真阳性率，即对真目标做出阳性反应的程度；而特异度对应于真阴性率，即对假目标做出阴性反应的程度，具体如下：

$$Sensitivity=\frac{TP}{TP+FN}\times 100\% \tag{47}$$

$$Specificity=\frac{TN}{FP+TN}\times 100\% \tag{48}$$

精确率（*Precision*）/召回率（*Recall*）/F1 分数（*F1*）. 在数据不均衡时[285]，准确率和错误率无法很好的评估分类模型，此时精确率、召回率、$F1$ 分数是更好的评价指标.

$$Precision=\frac{TP}{TP+FP} \tag{49}$$

$$Recall=\frac{TP}{TP+FN} \tag{50}$$

$$F1=\frac{2\times Precision\times Recall}{Precision+Recall} \tag{51}$$

精确率和召回率反映了模型对于特定类别的分辨能力，但是往往不可兼得：在提升一项的同时往往另一项会有所降低. 此时，可以使用 $F1$ 分数作为平衡精确率和召回率的评价指标. 在 $F1$ 指标较高时，精确率和召回率往往达到了较好的平衡.

PR 曲线/ROC 曲线. 在评价单标签分类模型时，若仅仅关注某一数值作为衡量指标，例如精确率、召回率等，通常不能准确反映分类器优劣情况，需要通过不同指标的综合分析来精确评价分类器模型性能，如 PR 曲线和 POC 曲线[286]等.

(1) PR 曲线(Precision-Recall curve)是以精确率作为纵坐标、召回率作为横坐标所绘制的曲线. PR 曲线可反映出分类器在类不平衡数据集下的分类性能,便于进行模型的优化和改进[287].

(2) ROC 曲线(Receiver Operating Characteristic curve)又称受试者工作特征曲线,是以真阳性率为纵坐标、假阳性率为横坐标所绘制的曲线. ROC 曲线适用于样本类分布较为均匀的情况,可对模型的泛化性能进行评估[287].

(3) *AUPRC*(Area Under the PR)表示 PR 曲线与 x 轴围成的面积,其数值近似于平均精度,通常适用于评估类分布不平衡时的场景[288].

(4) *AUROC*(Area Under the ROC)表示 ROC 曲线与 x 轴围成的面积[289]. *AUROC* 的数值介于 0～1 之间,当 *AUROC* 值越接近于 1 时,表示分类器可以较好地分类正负样本,性能较高;反之,则分类器性能较差.

精确匹配 *EM*(Exact Match). *EM* 是问答系统的一种常见的评价标准,它用来评价预测中精确匹配到正确答案(ground truth answers)的百分比[290].

平均排名 *MR*(Mean Rank). *MR* 是搜索算法通用的评价指标,在新闻分类、主题标记等文本分类任务中也有广泛应用. 其定义如下:

$$MR=\frac{1}{|Q|}\sum_{i=1}^{|Q|}|rank_i| \tag{52}$$

其中,Q 为查询数量,$rank_i$ 为第 i 个查询中第一个正确答案的排名.

在单标签文本分类任务中,该指标往往对应正确标签的平均排名. 当 *rank* 较为均衡时,*MR* 作为评价指标更为准确.

平均倒数排名 *MRR*(Mean Reciprocal Rank). *MRR* 是排名算法的常用评价指标,在问答系统以及信息检索中应用广泛[291],其定义如下:

$$MRR=\frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{|rank_i|} \tag{53}$$

其中,Q 为查询数量,$rank_i$ 为第 i 个查询中第一个正确答案的排名.

因为 *MRR* 只关注第一个正确答案的排名,故其更适合于只有一个正确答案的情况.

平均精度 *MAP*(Mean Average Precision). 在问题具有多个正确答案时,可以使用 *MAP* 对模型进行性能评估,其具体如下:

$$MAP=\frac{1}{|Q|}\sum_{i=1}^{Q}Ave(P(q)) \tag{54}$$

其中,Q 为查询数量,$P(q)$ 为对于 q 的查询精确分数[291].

其它广泛使用的指标包括错误发现率 *FDR*(False Discovery Rate)[292]和错误遗漏率 *FOR*(False Omission Rate)[293]等. 在实际场景中,通常需要在了解各指标具体反映了模型哪方面性能的基础上,再结合业务需求选择合适的评估指标.

在文本单标签多分类应用场景中,通常需要考虑各类别分布对结果的影响,可以采用如宏平均/微平均/加权平均等指标进行评价. 该场景与文本多标签分类场景类似,可以看作多标签分类问题的一种特殊情况,下文将在多标签分类场景中对此类评价指标进行详细介绍.

6.2 多标签评价指标

在许多应用场景中,样本通常与多个类别标签相关联,例如将一段新闻文本分为经济、文化等多个类别[294]. 每个样本与多个标签相关联,导致多标签分类任务中的性能评估比传统的单标签分类更为复杂. 由于任务目标存在本质差别,上文中为单标签文本分类所设计的评价指标不再适用,针对多标签文本分类模型需要重新设计评价指标用于性能评估.

宏平均/微平均/加权平均. 在处理多标签文本分类问题时,通常将该问题分解为多个二分类问题,每次以其中一个类为正类,其余类均为负类,计算之前提到的各类平均指标,最后再平均计算多标签评价指标,具体分为宏平均、微平均、加权平均三种方式[295].

(1) 微平均(*Micro-averaging*)首先计算总体类别的 TP,FN 和 FP 的值,再根据结果计算各类评价指标. 例如对于微平均精确率(*Mirco-precision*),其计算公式如下:

$$Mirco\text{-}precision=\frac{\sum_{i=1}^{|Y|}TP_i}{\sum_{i=1}^{|Y|}TP_i+\sum_{i=1}^{|Y|}FP_i} \tag{55}$$

(2) 宏平均(*Macro-averaging*)与微平均不同,宏平均先分别计算每个类别的评价指标,随后通过算术平均数计算得到最终平均结果. 例如对于宏平均召回率(*Marco-recall*),其计算公式如下:

$$Marco\text{-}recall=\frac{1}{|Y|}\sum_{i=1}^{|Y|}recall_i \tag{56}$$

(3) 加权平均(*Weighted-averaging*)是对宏平均的改进,考虑了每个类别的样本数量在总样本中的占比,解决了宏平均中所忽略的样本不均衡问题. 其核心思想是在计算平均评估指标时,各个类别的评

估指标要乘以该类在总样本所占的比例. 例如对于加权评价召回率(*Weighted-recall*)，其计算公式如下：

$$Weighted\text{-}recall=\frac{1}{|Y|}\sum_{i=1}^{|Y|}\frac{N_i}{N}\times recall_i \quad (57)$$

0-1 损失 *ZL*(Zero-one Loss). 0-1 损失是最为原始的损失函数，相对简单，多适用于分类问题. 若预测值与目标值不相等，那么为 1，否则为 0，其计算公式如下：

$$ZL(y_i,x_i)=\begin{cases}0, & y_i=x_i\\ 1, & y_i\neq x_i\end{cases} \quad (58)$$

其中 x_i 和 y_i 表示模型预测类别和样本真实类别.

绝对匹配率 *EMR*(Exact Match Ratio). 对于每一个样本而言，只有当预测类别与真实类别完全相同时才表示预测正确. 其计算公式如下：

$$EMR(N)=\frac{1}{N}\sum_{i=1}^{N}xnor(x_i,\ y_i) \quad (59)$$

其中 N 表示样本总数，x_i 和 y_i 分别表示模型预测类别和样本真实类别，$xnor$ 表示同或运算.

汉明损失 *HL*(Hamming Loss). *HL* 表示所有类别中错误预测样本的比例，所以该值越小则模型的分类能力越强[296]，其计算公式如下：

$$HL(N)=\frac{1}{N}\sum_{i=1}^{N}\frac{xor(x_i,y_i)}{|Y|} \quad (60)$$

其中 N 表示样本总数，$|Y|$ 表示标签类别总数，x_i 和 y_i 分别表示模型预测类别和样本真实类别，xor 表示异或运算.

除了单标签、多标签分类任务之外，文本分类中还存在着极端多标签分类任务. 极端多标签文本分类是指从一个非常大的标签集合中为每个文档分配其最相关的类别标签子集的问题，其中类别标签的数量可达数十万或数百万，巨大的标签空间带来了数据稀疏性和可扩展性等研究挑战[297]. 为了解决以上挑战，一些能够适用于极端多标签分类任务的基于排序的评价指标被提出.

前 *K* 精确率 *P@K*(*Precision* at Top *K*). *P@K* 是所有类别预测结果按相关性降序排序后，前 K 个结果中正确的预测类别所占比例[297]. 假设样本 t 的真实类别为 $\mathcal{L}_t=\{l_1,l_2,\cdots,l_{|\mathcal{L}|-1}\}$，根据预测概率降序排序后的类别列表为 $P_t=\{p_0,p_1,\cdots,p_{|Y|-1}\}$，其对应 *Precision@K* 计算公式如下：

$$P@K=\frac{1}{K}\sum_{i=1}^{K}rel(p_i,\mathcal{L}) \quad (61)$$

其中 rel 是相关性计算函数，用于判断样本预测类别是否与真实类别一致，其定义如下：

$$rel(p,\mathcal{L})=\begin{cases}1, & p\in\mathcal{L}\\ 0, & \text{其他}\end{cases} \quad (62)$$

累计增益 *CG*(Cumulative Gain). *CG* 通常应用于推荐系统模型评估，在多标签文本分类中也有应用. 在这种评估中，每个标签的相关性得分用作其在结果中排名位置的增益值度量，增益从排名位置 1 到 N 逐步相加[298]. 在多标签文本分类问题中，*CG* 反映了给出标签的得分总和，但不反应各个标签的位置因素. 若搜索结果列表有 M 个结果，rel_i 是第 i 位结果的增益值，其对应 *CG* 计算公式如下：

$$CG_M=\sum_{i=1}^{M}rel_i \quad (63)$$

折损累计增益 *DCG*(Discounted Cumulative Gain). *DCG* 作为 *CG* 评价方法的改进，考虑到了结果的位置因素带来的影响. 位置越靠后，文档的相关性得分加在累积增益中的份额越小[298]，具体计算公式如下：

$$DCG_M=rel_1+\sum_{i=2}^{M}\frac{rel_i}{\log_2 i} \quad (64)$$

前 *K* 归一化折损累计增益 *NDCG@K*(Normalized Discounted Cumulative Gain at Top *K*). *NDCG@K* 与 *DCG* 相同，除了考虑预测结果列表中类别标签的预测正确性之外，还对这些类别在列表中的位置因素所带来的影响进行建模[297]. *NDCG@K* 进一步地对不同查询的 *DCG* 值进行归一化便于比较，具体计算公式如下：

$$NDCG@K=\frac{1}{IDCG@K}\sum_{i=0}^{K}\frac{rel(p_i,\mathcal{L})}{\ln(i+1)} \quad (65)$$

其中，rel 是相关性计算函数，*IDCG@K* 是理想折损累积增益，计算公式如下：

$$IDCG@K=\sum_{j=0}^{\min(|\mathcal{L}|,K)}\frac{1}{\ln(j+1)} \quad (66)$$

总的来说，若各类别下样本规模分布均匀，则三个平均计算方法任选一个作为评价指标均可；若任务需求更加侧重于评估样本规模较大类别的预测效果，则应当选用微平均或者加权平均；若任务需求中各类别的预测效果重要程度一致，选用宏平均更为合适. 此外，当面临类别标签数量很大甚至极大这类极端情况，可以选择使用 *P@K* 或者 *NDCG@K*.

7　模型评估

基于上述模型与数据集，本文对不同种类模型在典型应用场景上的性能进行总结对比，如表 5 所示①. 文本分类模型类型主要包括传统机器学习模型、

① 表 5 中准确率结果均来自论文作者呈现结果.

表 5 不同文本分类模型在典型数据集下的准确率对比

模型类别	模型/方法	情感分析				新闻分类			抽取式问答	主题标记	自然语言推理
		MR	SST-2	IMDB	Yelp-2	20NG	R8	AG	TREC	DBpedia	MNLI
传统机器学习模型	AL-MSVM[77]	—	—	—	—	77.4	97.3	—	—	—	—
	Bayesian[66]	—	88.2	89.2	—	—	—	87.7	—	98.6	—
基础网络模型	FastText[50]	—	—	—	95.7	—	—	92.5	—	98.6	—
	TextCNN[105]	81.5	88.1	—	—	—	—	—	92.2	—	—
	Char-level CNN[134]	—	—	—	95.4	—	—	91.5	—	98.6	—
	C-LSTM[135]	—	—	—	—	—	—	—	**94.6**	—	—
胶囊网络模型	Capsule Network[93]	—	—	—	96.5	—	—	92.4	—	98.7	—
图神经网络模型	Text GCN[7]	76.7	—	—	—	86.3	97.1	—	—	—	—
	HGAT[97]	62.8	—	—	—	—	—	72.1	—	—	—
	HyperGAT[98]	78.3	—	—	—	**86.6**	**98.0**	—	—	—	—
迁移学习模型	ULMFiT[99]	—	—	—	97.8	—	—	95.0	—	99.2	—
预训练语言模型	BERT-base[30,250]	—	93.5	94.6	97.7	—	—	94.8	—	99.3	84.6
	BERT-large[30,250]	—	94.9	95.1	98.0	—	—	**95.1**	—	**99.4**	86.7
	GPT[110]	—	91.3	—	—	—	—	—	—	—	—
	SentiLARE[116]	**90.8**	—	**95.7**	**98.2**	—	—	—	—	—	—
	ELECTRA[120]	—	**97.1**	—	—	—	—	—	—	—	**91.3**
	ViLBERT[122,299]	—	90.4	—	—	—	—	—	—	—	80.1
	VisualBERT[123,299]	—	89.4	—	—	—	—	—	—	—	81.6
	BEIT-3[124]	—	92.6	—	—	—	—	—	—	—	83.8
提示学习模型	PET[113]	—	—	—	64.8	—	—	86.9	—	—	85.3
	KNN-Prompt[59]	78.2	84.2	—	—	—	—	78.8	—	—	—
	RLPrompt[115]	87.1	92.5	—	95.1	—	—	80.2	—	—	—
	DP_2O[186]	88.6	93.6	—	94.3	—	—	—	—	—	—

基础网络模型、胶囊网络模型、图神经网络模型等；典型数据集主要包括 MR、SST-2、IMDB、Yelp-2、20NG、R8、AG、TREC、DBpedia 和 MNLI，涵盖情感分析、新闻分类、抽取式问答、主题标记、自然语言推理等应用领域.

由于不同文本分类任务的特点，模型在各类任务中表现出显著的性能差异. 以情感分析为例，基于预训练语言模型的 SentiLARE[116] 在 Yelp-2 数据集上取得了较好的性能，准确率高达 98.2%. 优异的性能表现表明将句子情感间的联系融合到预训练任务中可以实现模型对文本信息的学习，有助于提高模型情感分析任务的表现. 在 SST-2 数据集上，基于预训练语言模型的 ELECTRA[120] 达到了 97.1% 的分类准确率，其表明利用对比学习思想可以使得预训练模型在使用较少参数的情况下也能学习到较高抽象层次的特征，从而实现较好的分类效果. 而在新闻分类任务中，HyperGAT[98] 作为图神经网络模型通过文档级超图来对文本进行建模. 这种建模方法更加贴合新闻在社交网络中的传播特性. 在 R8 数据集上，HyperGAT 实现了 98.0% 的最优分类准确率，而在 20NG 数据集上，该模型同样展现了 86.6% 的卓越性能. 在抽取式问答分类任务中，C-LSTM[135] 作为基础网络模型取得了良好性能，该方法组合了 CNN 和 LSTM 两种模型架构，结合不同模型架构的优势，在 TREC 数据集上取得了 94.6% 的准确率. 这一现象表明，通过对基础网络模型进行合理的设计，在全量数据训练的场景下，使用轻量级的模型结构依然有可能获得很好的文本分类效果.

通过综合不同应用场景下性能结果对比发现，相较于传统机器学习模型，深度学习模型在不同数据集上的表现更好，而对于深度学习模型，基于预训练语言模型的文本分类方法在大多数数据集上表现出卓越的性能. 例如，BERT[30,250] 应用于主题标记文本分类任务时，其在 DBpedia 数据集上的准确率高达 99.4%；在处理新闻分类任务时，该模型在 AG 数据集上也达到了 95.1% 的准确度. 此外，SentiLARE[116] 通过上下文感知的情感注意力机制和额外的标签感知掩码预训练，在情感分析相关的数据集上都达到了较高的准确度. 这表明预训练语言模型通过从海量未标记的数据集中学习到文本的潜在特征，对提高 NLP 任务的准确率起到了重要的作用. 总的来说，随着预训练模型技术的不断发展，基于预训练模型的分类方法在文本分类领域已取得了较为广泛的应用，且获得了较好的性能表现.

此外，以 PET[113]、RLPrompt[115] 为代表的基于提示学习的文本分类方法在数据集约束的实验场景下，依旧达到了很好的分类效果. 其中，RLPrompt[115] 方

法通过结合强化学习，制定了一个参数高效的策略网络以优化预训练模型的提示输入. 即使在无预训练模型梯度信息计算、每类仅有 16 个样本的低资源场景下，RLPrompt 也达到了与 BERT-large 等多种预训练语言模型在全量数据训练条件下相媲美的性能水平. 而 KNN-Prompt 方法[59]则更进一步，将文本分类推向了更加极端的零样本分类场景. 该方法通过利用 KNN 检索提示增强语言模型，显著提高了预训练模型在零样本条件下的分类效率. 与未经增强的 GPT-2 模型相比，KNN-Prompt 在性能上平均提升了 13.4%. 这些研究结果说明基于提示学习的方法可以通过优化提示，将不同的下游任务快速转换，较大限度地发掘预训练语言模型的潜力.

总体而言，目前文本分类领域的核心进展主要集中在深度学习和预训练语言模型的广泛应用上. 通过对特定任务优化提示输入或微调预训练模型，可以有效增强模型在各种场景中的文本分类性能. 随着相关技术的不断进步，文本分类领域预计也将持续快速发展. 本文对不同文本分类模型在不同种类数据集上的性能进行了全面深入的探讨，为研究者和从业者提供了重要的指导和启示. 这些研究结果对于选择、设计和优化文本分类模型具有实际应用价值，有助于推动文本分类领域的发展与创新.

8　文本分类挑战及展望

随着大数据时代的到来，深度学习在文本分类方面的应用已受到学术界和工业界的广泛研究，基于深度学习的文本分类模型已经成为当下的研究热点. 尽管一些新的文本分类模型反复刷新了大多数分类任务的准确率记录，但它们距达到像人类一样从语义层面“理解”文本仍有一段距离，在未来必将有更多、更广泛的尝试. 因此，本文分别从数据约束和模型计算两个角度进行总结，阐述当前研究所面临的挑战以及未来可能的研究方向.

8.1　数据约束方面

数据标注成本. 现有的大多数模型都是有监督模型，在训练过程中过度依赖标记数据，模型的性能将受到标注数据规模、质量、分布等要素的显著影响. 尽管近年来在常见的文本分类任务上涌现了大量大规模数据集，但是对于某些垂直领域内的应用（例如金融、农业和教育）以及更为复杂的文本分类任务（例如多模态数据、多语言和超长文档的文本分类），往往面临着标注数据缺乏的问题，这极大地限制了相关领域方法研究的发展[300-302]. 而新数据集的标注过程需要耗费大量的人力、物力以及时间，因此如何能够通过少量标注数据进行训练仍能取得良好效果的算法在未来有着很大的发展潜力，例如半监督学习和无监督学习文本分类算法[303].

同时，为了解决数据标注成本高的问题，人们将小样本/零次学习应用于文本分类，以实现对没有或者仅有少量相同类别训练数据的文本进行分类. 无监督学习通常只用未被标注的数据作为数据源，与无监督学习不同的是小样本/零次学习则是有效利用其它先验知识. 在人工智能趋近于人类的学术目标和廉价学习的工业需求推动下，小样本/零次学习研究引起了学术界与产业界的广泛关注[304]，是一个解决高标注成本的热门话题. 但是总体来说，目前应用于文本分类的研究还比较少，未来还待更深入和更广泛的研究.

外部知识利用. 通常情况下，对于特定机器学习分类模型，丰富的有效信息输入能够显著提升模型分类性能. 但是自然语言文本自身在各个层次上广泛存在歧义性和多义性等问题，给文本理解过程带来了极大的干扰[305]. 例如长文本或者超长文本的输入、阴阳怪气式语言蕴含的复杂语义、相同文本含义随着时间的演变以及特定领域内的专有名词、俚语和缩写等. 对上述问题的解决程度直接影响到模型的文本理解能力，进而影响模型性能. 除此之外，随着在线社交网络的发展，短文本分类任务也愈发常见，但是它们缺乏足够的上下文信息，这使得短文本分类面临着巨大的挑战.

有效引入外部知识是一个解决上述问题并提高模型性能的有效方法[306]，例如一个包含常识知识库的问答系统可以回答关于现实世界的问题，并且能够帮助解决信息不完整的难题. 现有的外部知识包括语义词典[307]、概念信息[308]、常识信息[309]、知识图谱[310]和多模态知识[311]等. 尽管外部知识的引入能够增强文本的语义表示，并已经在相关领域取得了一定成效[312]. 但是由于各知识特点的不同，在使用过程中会面临外部知识的选择以及如何有效建模和使用这类知识等问题，仍然需要进行更多的研究与实践.

多语种文本分类. 随着全球化的推进，多语种文本分类已经成为当前自然语言处理研究领域的一个重要课题. 然而，多语种文本之间存在语言差异，存在词汇、语法和表达方式不同等问题，这些都使得研

究人员难以利用现有的方法进行文本分类，因此需要提出适用于多语种文本的分类技术[313]. 除此之外，在多语种文本分类任务中目前并没有统一的评价标准，因此难以对不同方法的分类效果进行有效评估.

总体来说，多语种文本分类是一个需要综合考虑语言学、计算机技术和数据科学的研究领域. 研究人员必须考虑语言差异、数据不平衡、文本长度差异、词汇差异等问题，并对模型的泛化能力、标注问题、评价标准等进行全面评估，以提高多语种文本分类的效果.

多标签分类计算. 文本多标签分类任务中标签的规模大小、分布均衡程度和内在标签类别依赖性，都给该任务的解决带来了极大的挑战. 其中，在标签规模方面，极端多标签文本分类任务的难点在于标签集的数目非常多，包含数十万，甚至成百上千万的标签，而目前模型的内存占用、模型大小都随着标签空间的变大而线性变大，以致于难以部署甚至训练模型[314]；在标签分布均衡方面，部分数据集的标签对应的文本数据非常稀疏，存在“长尾”问题，这些不均衡的数据训练出来的模型会导致样本少的种类预测性能很差，甚至无法进行准确分类[315]；在标签的类别依赖方面，依赖性是指不同标签之间存在内在联系，比如属于“犯罪”的文本往往跟“法律”是相关联的，而现有的一些方法在处理多标签文本分类问题上，往往忽略了标签之间的依赖性，严重影响了模型的分类性能[316]. 因此，从数据层面来看，如何消除标签的规模、分布均衡性和类别依赖性所带来的影响是多标签文本分类任务未来的研究重点.

8.2 模型计算方面

跨模态文本学习. 高效的文本表示学习是文本分类模型的基础，相关研究伴随着语言学、认知科学和人工智能的发展而不断进步. 其核心方法从早期的规则法，到基于统计的方法，再到近年来的深度学习方法；表示对象从学习特定任务中字、词、句子、篇章的表示，再到端到端大规模自动文本表示学习. 总而言之，随着人工智能技术的发展、尤其是大规模预训练模型的出现，文本表示学习方法对文本中蕴含语义的理解程度越来越高. 基于 RNN 的模型、基于 CNN 的模型、基于注意力机制的模型以及基于 Transformer 的模型等方法理论基础成熟、模型简单有效，并且各种扩展均已取得不错的分类效果.

但是随着跨模态数据的出现，跨模态文本学习面临着新的挑战：(1) 多模态语义的融合表示学习方法研究，需要综合考虑声音、图像、视频、文本等不同模态的信息实现对信息的多维度和深层次理解；(2) 跨模态的统一信息表示方法研究，需要考虑将相同语义的不同模态的数据进行相同或相近的表示，试图为不同模态内容构建统一的数据表示模型，提高各模态内容的表示能力，以实现对现有表示模型的深入研究和广泛应用. 因此，如何基于已有研究基础[317]，结合跨模态文本数据表示学习新问题、新挑战，对这些方法进行深入研究、优化和融合，设计一个集成且资源依赖性小的模型范式[318]，有效解决跨模态文本学习难题，是未来一个重要研究方向.

模型的鲁棒性. 尽管现有的基于深度学习文本分类模型能够取得非常出色的分类成绩，但是随着应用场景越来越复杂，所使用的模型结构也越来越复杂，面临着训练难、参数多、推理慢等难题[317]. 为了提高模型的文本分类的准确性，研究人员往往在特定的数据集上，采用一定的训练技巧，使得模型达到更加优异的性能，但是随之而来的是模型对训练数据、模型参数等具有较强的依赖性.

此外，随着深度学习的对抗性攻击技术日益发展[319-320]，对抗性样本的危险性日益凸显，使得模型的脆弱性与鲁棒性问题研究日益重要. 因为这些深度伪造的样本通常不干扰人类的正常判断，但却能使机器学习模型做出错误判断、极大地降低模型准确率，给算法的安全性和鲁棒性带来了极大的挑战[321]. 目前，大多数方法采用对抗训练来提高模型的鲁棒性，如通过将对抗性样本加入训练过程，并取得了一定效果，但是方法的脆弱性难以抵抗日益复杂的攻击手段. 特别是在文本计算领域，大规模预训练模型与大规模训练数据集在文本分类的任务中扮演着不可替代的作用，而这些往往是对抗攻击容易得手的途径，会严重影响文本分类算法的性能. 因此，针对文本分类方法模型的鲁棒性研究，虽然取得一些成绩，但是仍需要进行更加深入的探讨和研究.

模型的可解释性. 尽管对于深度学习方法，复杂的非线性计算赋予了它们极高的模型表示能力，但是其本质上都是在进行曲线拟合过程. 最后只能得到一堆看上去毫无意义的模型参数和拟合度非常高的判定结果，其中蕴含的知识难以用人们所能理解的方式进行表达[322]. 如果一个模型完全不可解释，那么在许多领域的应用会因为没办法得到更多可靠的信息而受到限制. 例如情感分析系统中，除了最终的情感判定结果外，人们还需要了解模型产生这样的判定是基于文本中哪些因素的考虑，以便利用人

工交互信息提升现有方法性能.此外,可解释性对模型优化也有一定影响,例如不明确模型究竟能够学习到哪些特征以及模型在不同数据集有着不同表现的原因等问题,会导致失去了对模型的改进和优化的指导方向.因此,有必要对模型的可解释性进行深入研究,能够兼顾效率、准确度、可解释性的模型将在很多应用场景中具备不可替代的优势.

应用大语言模型的文本分类. 随着大语言模型(例如 GPT-3、GPT-4[323]等)的出现和广泛应用,文本分类计算范式正在经历一场革命性的演变.与以往需要大量标签数据进行训练的模型不同,这些大语言模型凭借其巨大的模型参数量和海量非标注文本数据上的预训练,使其在处理小样本或者零样本文本分类问题上表现出色.

但与此同时,大语言模型的出现也引入了新的挑战.首先,数十亿甚至数百亿的参数使得这些模型需要大量的计算资源,大幅增加了模型训练和部署的成本[324].因此如何轻量化模型以适应更多场景(例如,终端高实时性需求场景)将是未来一个重要的研究方向;其次,由于目前的大语言模型往往作为黑箱模型运行,缺乏透明性和可解释性,导致在某些文本分类的应用场景(例如医疗、法律)缺乏证据或者事实支撑而无法使用[325];此外,由于这些模型通常会使用全网数据进行训练,在处理文本分类任务时大语言模型可能会生成不准确或具有偏见的内容表征[326-327].因此,如何对繁杂的网络数据进行有效的预处理,以减少数据偏见并提高大语言模型的公平性也是值得关注的问题.

面向大规模数据的文本分类. 随着大规模数据在文本分类领域的普遍使用,不仅面临着数据的存储、计算和安全等问题,同时在模型层面也迎来了许多新的挑战.例如大规模数据带来的维数灾难等问题大大增加了模型的训练难度;同时随着数据规模的不断扩大,模型所需要的模型结构也越来越复杂,面临着参数多、推理慢、泛化能力弱等缺点;并且在大规模数据的文本分类中,评价模型性能的方法往往不够全面,因此很难评价模型的真实效果.这些问题大大限制了模型在高实时性需求场景的线上应用能力[317].因此,如何设计更有效率、高精度的文本分类技术,同时提高模型的收敛性和在线推理速度,有效地解决面向大规模数据的文本分类难题,为相关技术研究创造更为广阔的应用空间,是文本分类技术得以发展的一个重要研究方向.

9　总　结

本文对文本分类领域从传统机器学习方法到新兴深度学习方法的研究成果进行了详尽的调研.针对传统文本分类方法,从常见的基于传统机器学习方法的分类器原理方面进行了详细分析,包括概率图模型、支持向量机模型和决策树模型等.同时为了能让读者针对不同的应用场景选择合适的分类模型,本文详尽地介绍了不同分类模型在处理各类文本分类任务上的优势和不足.对基于深度学习的文本分类方法,本文根据其发展脉络与计算核心思想的不同进行分类阐述,其中包括基础网络模型、胶囊网络模型、图神经网络模型、迁移学习模型、预训练模型等.此外,本文还对不同应用任务场景下常用的数据集和模型评价方法进行了详细介绍,并以此为基础对不同种类文本分类模型在典型应用场景中进行了性能对比.同时,在分析目前领域内所存在研究挑战的基础上,本文展望了多个未来重要的研究方向,包括但不限于数据约束、模型计算等领域,力求为未来的研究者提供一个全面而深入的指导框架.本文相关工作的综述可以帮助读者快速理清文本分类技术的发展脉络,了解文本分类技术的国内外研究现状,有助于读者在大量的相关研究文献中快速理解问题背景,希望能够对相关领域的研究工作和工程技术人员提供有益的帮助.

参　考　文　献

[1] Tai K S, Socher R, Manning C D. Improved semantic representations from tree-structured long short-term memory networks //Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL 2015) and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing. Volume 1: Long Papers. Beijing, China, 2015: 1556-1566

[2] Zhu X, Sobhani P, Guo H. Long short-term memory over recursive structures//Proceedings of the 32nd International Conference on Machine Learning (ICML 2015). JMLR Workshop and Conference Proceedings: Volume 37. Lille, France, 2015: 1604-1612

[3] Chen J, Gong Z, Liu W. A Dirichlet process biterm-based mixture model for short text stream clustering. Applied Intelligence, 2020, 50(5): 1609-1619

[4] Chen J, Gong Z, Liu W. A nonparametric model for online topic discovery with word embeddings. Information Sciences, 2019, 504: 32-47

[5] Kalchbrenner N, Grefenstette E, Blunsom P. A convolutional neural network for modelling sentences//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics(ACL 2014). Volume 1: Long Papers. Baltimore, USA, 2014: 655- 665

[6] Liu P, Qiu X, Chen X, et al. Multi-timescale long short-term memory neural network for modelling sentences and documents//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015). Lisbon, Portugal, 2015: 2326- 2335

[7] Yao L, Mao C, Luo Y. Graph convolutional networks for text classification//Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI 2019), the 31st Innovative Applications of Artificial Intelligence Conference(IAAI 2019), the 9th AAAI Symposium on Educational Advances in Artificial Intelligence(EAAI 2019). Honolulu, USA, 2019: 7370-7377

[8] Johnson R, Zhang T. Deep pyramid convolutional neural networks for text categorization//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics(ACL 2017). Volume 1: Long Papers. Vancouver, Canada, 2017: 562-570

[9] Chen C, Zhang X, Ju S, et al. AntProphet: An intention mining system behind Alipay's intelligent customer service BOT//Proceedings of the 28th International Joint Conference on Artificial Intelligence(IJCAI 2019). Macao, China, 2019: 6497-6499

[10] Pan L, Zhang J, Lee P P C, et al. An intelligent customer care assistant system for large-scale cellular network diagnosis //Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Halifax, Canada, 2017: 1951-1959

[11] Chaudhuri S, Dayal U, Narasayya V. An overview of business intelligence technology. Communications of the ACM, 2011, 54(8): 88-98

[12] Godbole S, Roy S. Text classification, business intelligence, and interactivity: Automating C-Sat analysis for services industry //Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Las Vegas, USA, 2008: 911- 919

[13] Badaro G, Baly R, Hajj H, et al. A survey of opinion mining in Arabic: A comprehensive system perspective covering challenges and advances in tools, resources, models, applications, and visualizations. ACM Transactions on Asian and Low-Resource Language Information Processing, 2019, 18(3): 1-52

[14] Miao Z, Li Y, Wang X, et al. Snippext: Semi-supervised opinion mining with augmented data//Proceedings of the Web Conference 2020(WWW'20). Taipei, China, 2020: 617-628

[15] Apté C, Damerau F, Weiss S M. Automated learning of decision rules for text categorization. ACM Transactions on Information Systems, 1994, 12(3): 233-251

[16] Lewis D D, Ringuette M. A comparison of two learning algorithms for text categorization//Proceedings of the 3rd Annual Symposium on Document Analysis and Information Retrieval. Volume 33. Las Vegas, USA, 1994: 81-93

[17] Scott S, Matwin S. Feature engineering for text classification //Proceedings of the 16th International Conference on Machine Learning(ICML 1999). Bled, Slovenia, 1999: 379-388

[18] Sebastiani F. Machine learning in automated text categorization. ACM Computing Surveys, 2002, 34(1): 1-47

[19] Zhang Y, Jin R, Zhou Z H. Understanding bag-of-words model: A statistical framework. International Journal of Machine Learning and Cybernetics, 2010, 1(1): 43-52

[20] Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. Information Processing & Management, 1988, 24(5): 513-523

[21] Maron M E. Automatic indexing: An experimental inquiry. Journal of the ACM, 1961, 8(3): 404-417

[22] Altman N S. An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician, USA, 1992, 46(3): 175-185

[23] Joachims T. Text categorization with support vector machines: Learning with many relevant features//Proceedings of the 10th European Conference on Machine Learning. Lecture Notes in Computer Science: Volume 1398 Machine Learning: ECML-98. Chemnitz, Germany, 1998: 137-142

[24] Bengio Y, Ducharme R, Vincent P. A neural probabilistic language model//Advances in Neural Information Processing Systems 13, Papers from Neural Information Processing Systems (NIPS) 2000. Denver, USA, 2000: 932-938

[25] Lecun Y, Bengio Y, Hinton G. Deep learning. Nature, 2015, 521(7553): 436-444

[26] Ghosal D, Akhtar M S, Chauhan D S, et al. Contextual inter-modal attention for multi-modal sentiment analysis// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium, 2018: 3454-3466

[27] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality// Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Lake Tahoe, USA, 2013: 3111-3119

[28] Ilic S, Marrese-Taylor E, Balazs J A, et al. Deep contextualized word representations for detecting sarcasm and irony// Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA@EMNLP 2018). Brussels, Belgium, 2018: 2-7

[29] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need//Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017. Long Beach, USA, 2017: 5998-6008

[30] Devlin J, Chang M, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019). Volume 1 (Long and Short Papers). Minneapolis, USA, 2019: 4171-4186

[31] Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners. OpenAI Blog, 2019, 1(8): 9

[32] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. Advances in Neural Information Processing Systems, 2020, 33: 1877-1901

[33] Li Q, Peng H, Li J, et al. A survey on text classification: From traditional to deep learning. ACM Transactions on Intelligent Systems and Technology, 2022, 13(2): 1-41

[34] Minaee S, Kalchbrenner N, Cambria E, et al. Deep learning-based text classification: A comprehensive review. ACM Computing Surveys, 2021, 54(3): 1-40

[35] Kowsari K, Jafari Meimandi K, Heidarysafa M, et al. Text classification algorithms: A survey. Information, 2019, 10(4): 150

[36] Aggarwal C C, Zhai C. A survey of text classification algorithms//Aggarwal C C, Zhai C. Mining Text Data. New York: Springer, 2012: 163-222

[37] Wu T, Caccia M, Li Z, et al. Pretrained language model in continual learning: A comparative study//Proceedings of the 10th International Conference on Learning Representations (ICLR 2022). Virtual Event, 2022

[38] Zhuang F, Qi Z, Duan K, et al. A comprehensive survey on transfer learning. Proceedings of the IEEE, 2020, 109(1): 43-76

[39] Liu P, Yuan W, Fu J, et al. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. ACM Computing Surveys, 2023, 55(9): 1-35

[40] Su Jin-Shu, Zhang Bo-Feng, Xu Xin. Advances in machine learning based text categorization. Journal of Software, 2006, 17(9): 1848-1859(in Chinese)
(苏金树，张博锋，徐昕. 基于机器学习的文本分类技术研究进展. 软件学报，2006，17(9)：1848-1859)

[41] Pang Liang, Lan Yan-Yan, Xu Jun, et al. A survey on deep text matching. Chinese Journal of Computers, 2017, 40(4): 985-1003(in Chinese)
(庞亮，兰艳艳，徐君等. 深度文本匹配综述. 计算机学报，2017，40(4)：985-1003)

[42] Xue Chun-Xiang, Zhang Yu-Fang. Research review on Chinese text classification in the news field. Library and Information Service, 2013, 57(14): 134-139(in Chinese)
(薛春香，张玉芳. 面向新闻领域的中文文本分类研究综述. 图书情报工作，2013，57(14)：134-139)

[43] Gupta G, Malhotra S. Text document tokenization for word frequency count using rapid miner//Proceedings of the International Conference on Advancements in Engineering and Technology (ICAET 2015). New York, USA, 2015: 24-26

[44] Verma T, Renu R, Gaur D. Tokenization and filtering process in RapidMiner. International Journal of Applied Information Systems, 2014, 7(2): 16-18

[45] Saif H, Fernández M, He Y, et al. On stopwords, filtering and data sparsity for sentiment analysis of Twitter//Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014). Reykjavik, Iceland, 2014: 810-817

[46] Mawardi V C. Spelling correction for text documents in Bahasa Indonesia using finite state automata and Levinshtein distance method. MATEC Web of Conferences, 2018, 164: 01047

[47] Hong Y, Yu X, He N, et al. FASPell: A fast, adaptable, simple, powerful Chinese spell checker based on DAE-decoder paradigm//Proceedings of the 5th Workshop on Noisy User-generated Text (WNUT@EMNLP 2019). Hong Kong, China, 2019: 160-169

[48] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space//Proceedings of the 1st International Conference on Learning Representations (ICLR 2013). Scottsdale, USA, 2013: 10-23

[49] Pennington J, Socher R, Manning C D. GloVe: Global vectors for word representation//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014). Doha, Qatar, 2014: 1532-1543

[50] Joulin A, Grave E, Bojanowski P, et al. Bag of tricks for efficient text classification//Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017). Volume 2: Short Papers. Valencia, Spain, 2017: 427-431

[51] Melamud O, Goldberger J, Dagan I. Context2vec: Learning generic context embedding with bidirectional LSTM//Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning (CoNLL 2016). Berlin, Germany, 2016: 51-61

[52] Uysal A K, Gunal S. The impact of preprocessing on text classification. Information Processing & Management, 2014, 50(1): 104-112

[53] Cover T M, Hart P E. Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 1953, 13(1): 21-27

[54] Jiang S, Pang G, Wu M, et al. An improved K-nearest-neighbor algorithm for text categorization. Expert Systems with Applications, 2012, 39(1): 1503-1509

[55] Abu Alfeilat H A, Hassanat A B, Lasassmeh O, et al. Effects of distance measure choice on K-nearest neighbor classifier performance: A review. Big Data, 2019, 7(4): 221-248

[56] Liu L, Zhou T, Long G, et al. Many-class few-shot learning on multi-granularity class hierarchy. IEEE Transactions on Knowledge and Data Engineering, 2022, 34(5): 2293-2305

[57] Isnain A R, Supriyanto J, Kharisma M P. Implementation of K-Nearest Neighbor (K-NN) algorithm for public sentiment analysis of online learning. Indonesian Journal of Computing and Cybernetics Systems, 2021, 15(2): 121-130

[58] Kamalloo E, Rezagholizadeh M, Passban P, et al. Not far away, not so close: Sample efficient nearest neighbour data augmentation via MiniMax//Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021. 2021: 3522-3533

[59] Shi W, Michael J, Gururangan S, et al. Nearest neighbor zero-shot inference//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP

2022). Abu Dhabi, United Arab Emirates, 2022: 3254-3265

[60] AG Corpus. http://www.di.unipi.it/~gulli/AG_corpus_of_news_articles.html, 2004

[61] YahooA Corpus. https://www.kaggle.com/datasets/soumikrakshit/yahoo-answers-dataset, 2015

[62] SST Corpus. http://nlp.stanford.edu/sentiment, 2013

[63] Dadgar S M H, Araghi M S, Farahani M M. A novel text mining approach based on TF-IDF and support vector machine for news classification//Proceedings of the 2016 IEEE International Conference on Engineering and Technology (ICETECH). Coimbatore, India, 2016: 112-116

[64] Londo G L Y, Kartawijaya D H, Ivariyani H T, et al. A study of text classification for Indonesian news article//Proceedings of the 2019 International Conference of Artificial Intelligence and Information Technology (ICAIIT). Yogyakarta, Indonesia, 2019: 205-208

[65] Kang M, Ahn J, Lee K. Opinion mining using ensemble text hidden Markov models for text classification. Expert Systems with Applications, 2018, 94: 218-227

[66] Mukherjee S, Awadallah A. Uncertainty-aware self-training for few-shot text classification. Advances in Neural Information Processing Systems, 2020, 33: 21199-21212

[67] Bahassine S, Madani A, Kissi M. An improved Chi-sqaure feature selection for Arabic text classification using decision tree//Proceedings of the 11th International Conference on Intelligent Systems: Theories and Applications (SITA 2016). Mohammedia, Morocco, 2016: 1-5

[68] Taloba A I, Ismail S S. An intelligent hybrid technique of decision tree and genetic algorithm for E-mail spam detection//Proceedings of the 2019 9th International Conference on Intelligent Computing and Information Systems (ICICIS). Cairo, Egypt, 2019: 99-104

[69] Yuvaraj N, Chang V, Gobinathan B, et al. Automatic detection of cyberbullying using multi-feature based artificial intelligence with deep decision tree classification. Computers & Electrical Engineering, 2021, 92: 107186

[70] Shi L, Ma X, Xi L, et al. Rough set and ensemble learning based semi-supervised algorithm for text classification. Expert Systems with Applications, 2011, 38(5): 6300-6306

[71] Kilinç D. The effect of ensemble learning models on Turkish text classification. Celal Bayar University Journal of Science, 2016, 12(2): 215-220

[72] Al-Ash H S, Putri M F, Mursanto P, et al. Ensemble learning approach on Indonesian fake news classification//Proceedings of the 2019 3rd International Conference on Informatics and Computational Sciences (ICICoS). Semarang, 2019: 1-6

[73] Cortes C, Vapnik V. Support-vector networks. Machine Learning, 1995, 20: 273-297

[74] Wang Z Q, Sun X, Zhang D X, et al. An optimal SVM-based text classification algorithm//Proceedings of the 2006 International Conference on Machine Learning and Cybernetics. Dalian, China, 2006: 1378-1381

[75] Wan C H, Lee L H, Rajkumar R, et al. A hybrid text classification approach with low dependency on parameter by integrating *K*-nearest neighbor and support vector machine. Expert Systems with Applications, 2012, 39(15): 11880-11888

[76] Garla V, Taylor C, Brandt C. Semi-supervised clinical text classification with Laplacian SVMs: An application to cancer case management. Journal of Biomedical Informatics, 2013, 46(5): 869-875

[77] Goudjil M, Koudil M, Bedda M, et al. A novel active learning method using SVM for text classification. International Journal of Automation and Computing, 2018, 15(3): 290-298

[78] BBC Datasets. http://nlp.stanford.edu/sentiment, 2013

[79] 20NG Corpus. http://ana.cachopo.org/datasets-for-single-label-text-categorization, 2007

[80] Xu S. Bayesian Naïve Bayes classifiers to text classification. Journal of Information Science, 2018, 44(1): 48-59

[81] Friedman N. Inferring cellular networks using probabilistic graphical models. Science, 2004, 303(5659): 799-805

[82] Ebrahimi J, Dou D, Lowd D. Weakly supervised tweet stance classification by relational bootstrapping//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016). Austin, USA, 2016: 1012-1017

[83] Socher R, Perelygin A, Wu J, et al. Recursive deep models for semantic compositionality over a sentiment Treebank//Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013). Seattle, USA, 2013: 1631-1642

[84] Safavian S R, Landgrebe D. A survey of decision tree classifier methodology. IEEE Transactions on Systems, Man, and Cybernetics, 1991, 21(3): 660-674

[85] Vens C, Struyf J, Schietgat L, et al. Decision trees for hierarchical multi-label classification. Machine Learning, 2008, 73: 185-214

[86] Friedman J H. Greedy function approximation: A gradient boosting machine. Annals of Statistics, 2001, 29: 1189-1232

[87] Ho T K. Random decision forests//Proceedings of the 3rd International Conference on Document Analysis and Recognition (ICDAR 1995). Volume I. Montreal, Canada, 1995: 278-282

[88] Chatzimparmpas A, Martins R M, Kucher K, et al. StackGenVis: Alignment of data, algorithms, and models for stacking ensemble learning using performance metrics. IEEE Transactions on Visualization and Computer Graphics, 2020, 27(2): 1547-1557

[89] Zeng K, Pan Z, Xu Y, et al. An ensemble learning strategy for eligibility criteria text classification for clinical trial recruitment: Algorithm development and validation. JMIR Medical Informatics, 2020, 8(7): e17832

[90] Silver D, Huang A, Maddison C J, et al. Mastering the game of go with deep neural networks and tree search. Nature, 2016, 529(7587): 484-489

[91] Yang M, Zhao W, Ye J, et al. Investigating capsule networks with dynamic routing for text classification//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium, 2018: 3110-3119

[92] Kim J, Jang S, Park E, et al. Text classification using capsules. Neurocomputing, 2020, 376: 214-221

[93] Ren H, Lu H. Compositional coding capsule network with *K*-means routing for text classification. Pattern Recognition Letters, 2022, 160: 1-8

[94] Peng H, Li J, He Y, et al. Large-scale hierarchical text classification with recursively regularized deep graph-CNN// Proceedings of the 2018 World Wide Web Conference on World Wide Web (WWW 2018). Lyon, France, 2018: 1063-1072

[95] Zhang C, Li Q, Song D. Aspect-based sentiment classification with aspect-specific graph convolutional networks// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019). Hong Kong, China, 2019: 4567-4577

[96] Zhang Y, Yu X, Cui Z, et al. Every document owns its structure: Inductive text classification via graph neural networks//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020). Online, 2020: 334-339

[97] Hu L, Yang T, Shi C, et al. Heterogeneous graph attention networks for semi-supervised short text classification// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019). Hong Kong, China, 2019: 4820-4829

[98] Ding K, Wang J, Li J, et al. Be more with less: Hypergraph attention networks for inductive text classification// Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020). Online, 2020: 4927-4936

[99] Howard J, Ruder S. Universal language model fine-tuning for text classification//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018). Volume 1: Long Papers. Melbourne, Australia, 2018: 328-339

[100] Banerjee S, Akkaya C, Perez-Sorrosal F, et al. Hierarchical transfer learning for multi-label text classification//Proceedings of the 57th Conference of the Association for Computational Linguistics (ACL 2019). Volume 1: Long Papers. Florence, Italy, 2019: 6295-6300

[101] Cao Z, Zhou Y, Yang A, et al. Deep transfer learning mechanism for fine-grained cross-domain sentiment classification. Connection Science, 2021, 33(4): 911-928

[102] Unanue I J, Haffari G, Piccardi M. T3L: Translate-and-test transfer learning for cross-lingual text classification. arXiv preprint arXiv:2306.04996, 2023

[103] Deng Wan-Yu, Zheng Qing-Hua, Chen Lin, et al. Research on extreme learning of neural networks. Chinese Journal of Computers, 2010, 33(2): 279-287(in Chinese)
(邓万宇，郑庆华，陈琳等. 神经网络极速学习方法研究. 计算机学报，2010，33(2)：279-287)

[104] Zhou Hang-Xia, Ye Jia-Jun, Ren Hua. Text classification based on fast auto-encoder RELM. Computer Engineering & Science, 2016, 38(5): 871-876(in Chinese)
(周杭霞，叶佳骏，任欢. 基于快速自编码的 RELM 的文本分类. 计算机工程与科学，2016，38(5)：871-876)

[105] Chen Y. Convolutional Neural Network for Sentence Classification[M. S. dissertation]. University of Waterloo, Waterloo, Canada, 2015

[106] Nguyen H, Nguyen M. A deep neural architecture for sentence-level sentiment classification in Twitter social networking//Communications in Computer and Information Science: Volume 781 Computational Linguistics-15th International Conference of the Pacific Association for Computational Linguistics (PACLING 2017). Yangon, Myanmar, 2017: 15-27

[107] Basiri M E, Nemati S, Abdar M, et al. ABCDM: An attention-based bidirectional CNN-RNN deep model for sentiment analysis. Future Generation Computer Systems, 2021, 115: 279-294

[108] Song X, Petrak J, Roberts A. A deep neural network sentence level classification method with context information //Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium, 2018: 900-904

[109] Sun S, Sun Q, Zhou K, et al. Hierarchical attention prototypical networks for few-shot text classification//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019). Hong Kong, China, 2019: 476-485

[110] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training. 2018

[111] Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback//Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022 (NeurIPS 2022). New Orleans, USA, 2022

[112] Yang Z, Liu Y. On robust prefix-tuning for text classification //Proceedings of the 10th International Conference on Learning Representations (ICLR 2022). Virtual Event, 2022

[113] Schick T, Schütze H. Exploiting cloze-questions for few-shot text classification and natural language inference//Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume (EACL 2021). Online, 2021: 255-269

[114] Liu X, Zheng Y, Du Z, et al. GPT understands, too. arXiv preprint arXiv:2103.10385, 2021

[115] Deng M, Wang J, Hsieh C, et al. RLPrompt: Optimizing discrete text prompts with reinforcement learning//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP 2022). Abu Dhabi, Emirates, 2022: 3369-3391

[116] Ke P, Ji H, Liu S, et al. SentiLARE: Sentiment-aware language representation learning with linguistic knowledge// Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020). Online, 2020: 6975-6988

[117] Araci D. FinBERT: Financial sentiment analysis with pre-trained language models. arXiv preprint arXiv:1908.10063, 2019

[118] Liu W, Zhou P, Zhao Z, et al. K-BERT: Enabling language representation with knowledge graph//Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI 2020), the 32nd Innovative Applications of Artificial Intelligence Conference (IAAI 2020), the 10th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI 2020). New York, USA, 2020: 2901-2908

[119] Sun Y, Wang S, Feng S, et al. ERNIE 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. arXiv preprint arXiv:2107.02137, 2021

[120] Clark K, Luong M, Le Q V, et al. ELECTRA: Pre-training text encoders as discriminators rather than generators//Proceedings of the 8th International Conference on Learning Representations (ICLR 2020). Addis Ababa, Ethiopia, 2020

[121] Xiong W, Du J, Wang W Y, et al. Pretrained encyclopedia: Weakly supervised knowledge-pretrained language model//Proceedings of the 8th International Conference on Learning Representations (ICLR 2020). Addis Ababa, Ethiopia, 2020

[122] Lu J, Batra D, Parikh D, et al. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks//Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019 (NeurIPS 2019). Vancouver, Canada, 2019: 13-23

[123] Li L H, Yatskar M, Yin D, et al. VisualBERT: A simple and performant baseline for vision and language. arXiv preprint arXiv:1908.03557, 2019

[124] Wang W, Bao H, Dong L, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. arXiv preprint arXiv:2208.10442, 2022

[125] Bebis G, Georgiopoulos M. Feed-forward neural networks. IEEE Potentials, 1994, 13(4): 27-31

[126] Pascanu R, Mikolov T, Bengio Y. On the difficulty of training recurrent neural networks//Proceedings of the 30th International Conference on Machine Learning (ICML 2013). JMLR Workshop and Conference Proceedings: Volume 28. Atlanta, USA, 2013: 1310-1318

[127] Lecun Y, Bengio Y, et al. Convolutional networks for images, speech, and time series. The Handbook of Brain Theory and Neural Networks, 1995, 3361(10): 255-258

[128] Mikolov T, Karafiát M, Burget L, et al. Recurrent neural network based language model//Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH 2010). Makuhari, Japan, 2010: 1045-1048

[129] Tang D, Qin B, Liu T. Document modeling with gated recurrent neural network for sentiment classification//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015). Lisbon, Portugal, 2015: 1422-1432

[130] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735-1780

[131] Zhang T, Huang M, Zhao L. Learning structured representation for text classification via reinforcement learning//Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18). New Orleans, USA, 2018: 6053-6060

[132] Kaelbling L P, Littman M L, Moore A W. Reinforcement learning: A survey. Journal of Artificial Intelligence Research, 1996, 4: 237-285

[133] Zhou Fei-Yan, Jin Lin-Peng, Dong Jun, et al. Review of convolutional neural network. Chinese Journal of Computers, 2017, 40(6): 1229-1251(in Chinese)
(周飞燕，金林鹏，董军等. 卷积神经网络研究综述. 计算机学报，2017, 40(6): 1229-1251)

[134] Zhang X, Zhao J, Lecun Y. Character-level convolutional networks for text classification//Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015. Montreal, Canada, 2015: 649-657

[135] Zhou C, Sun C, Liu Z, et al. A C-LSTM neural network for text classification. arXiv preprint arXiv:1511.08630, 2015

[136] TREC Corpus. https://cogcomp.seas.upenn.edu/Data/QA/QC/, 2002

[137] Zhang S, Jiang H, Xu M, et al. The fixed-size ordinally-forgetting encoding method for neural network language models//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL 2015). Volume 2: Short Papers. Beijing, China, 2015: 495-500

[138] Ni J, Li J, Mcauley J J. Justifying recommendations using distantly-labeled reviews and fine-grained aspects//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019). Hong Kong, China, 2019: 188-197

[139] Chen H, Ma Q, Lin Z, et al. Hierarchy-aware label semantics matching network for hierarchical text classification//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL/IJCNLP 2021). Volume 1: Long Papers. Virtual Event, 2021: 4370-4379

[140] Lewis D D, Yang Y, Russell-Rose T, et al. RCV1: A new benchmark collection for text categorization research. Journal of Machine Learning Research, 2004, 5(Apr): 361-397

[141] Sabour S, Frosst N, Hinton G E. Dynamic routing between capsules//Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017. Long Beach, USA, 2017: 3856-3866

[142] Xi E, Bing S, Jin Y. Capsule network performance on complex data. arXiv preprint arXiv:1712.03480, 2017

[143] Patrick M K, Adekoya A F, Mighty A A, et al. Capsule networks — A survey. Journal of King Saud University-Computer and Information Sciences, 2022, 34(1): 1295-1310

[144] Zhao F, Wu Z, He L, et al. Label-correction capsule network for hierarchical text classification. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023, 31: 2158-2168

[145] DBpedia Corpus. https://www.kaggle.com/datasets/danofer/dbpedia-classes, 2015

[146] Scarselli F, Gori M, Tsoi A C, et al. The graph neural network model. IEEE Transactions on Neural Networks, 2008, 20(1): 61-80

[147] Xu Bing-Bing, Cen Ke-Ting, Huang Jun-Jie, et al. A survey on graph convolutional neural network. Chinese Journal of Computers, 2020, 43(5): 755-780(in Chinese)
(徐冰冰，岑科廷，黄俊杰等. 图卷积神经网络综述. 计算机学报, 2020, 43(5): 755-780)

[148] Bastings J, Titov I, Aziz W, et al. Graph convolutional encoders for syntax-aware neural machine translation//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing(EMNLP 2017). Copenhagen, Denmark, 2017: 1957-1967

[149] Li Y, Jin R, Luo Y. Classifying relations in clinical narratives using segment graph convolutional and recurrent neural networks (Seg-GCRNs). Journal of the American Medical Informatics Association, 2019, 26(3): 262-268

[150] Li Z, Cai J, He S, et al. Seq2seq dependency parsing//Proceedings of the 27th International Conference on Computational Linguistics(COLING 2018). Santa Fe, USA, 2018: 3203-3214

[151] Xu K, Hu W, Leskovec J, et al. How powerful are graph neural networks?//Proceedings of the 7th International Conference on Learning Representations(ICLR 2019). New Orleans, USA, 2019

[152] Mihalcea R, Tarau P. TextRank: Bringing order into text//Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing(EMNLP 2004), A meeting of SIGDAT, a Special Interest Group of the ACL, Held in Conjunction with ACL 2004. Barcelona, Spain, 2004: 404-411

[153] Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks//Proceedings of the 5th International Conference on Learning Representations(ICLR 2017). Toulon, France, 2017: 132-146

[154] Defferrard M, Bresson X, Vandergheynst P. Convolutional neural networks on graphs with fast localized spectral filtering//Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016. Barcelona, Spain, 2016: 3837-3845

[155] Reuters Corpus. https://www.cs.umb.edu/~smimarog/textmining/datasets/, 2007

[156] Liu R, Shi Y, Ji C, et al. A survey of sentiment analysis based on transfer learning. IEEE Access, 2019, 7: 85401-85412

[157] Hussain M, Bird J J, Faria D R. A study on CNN transfer learning for image classification//Advances in Computational Intelligence Systems: Contributions Presented at the 18th UK Workshop on Computational Intelligence. Nottingham, UK, 2019: 191-202

[158] Morid M A, Borjali A, Del Fiol G. A scoping review of transfer learning research on medical image analysis using ImageNet. Computers in Biology and Medicine, 2021, 128: 104115

[159] Chen S, Ma K, Zheng Y. Med3D: Transfer learning for 3D medical image analysis. arXiv preprint arXiv:1904.00625, 2019

[160] Houlsby N, Giurgiu A, Jastrzebski S, et al. Parameter-efficient transfer learning for NLP//Proceedings of the 36th International Conference on Machine Learning(ICML 2019). Proceedings of Machine Learning Research: Volume 97. Long Beach, USA, 2019: 2790-2799

[161] Raffel C, Shazeer N, Roberts A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research, 2020, 21(1): 5485-5551

[162] NLI Corpus. https://cims.nyu.edu/~sbowman/xnli/, 2018

[163] Qiu X, Sun T, Xu Y, et al. Pre-trained models for natural language processing: A survey. Science China Technological Sciences, 2020, 63(10): 1872-1897

[164] Jang B, Kim M, Harerimana G, et al. Bi-LSTM model to increase accuracy in text classification: Combining word2vec CNN and attention mechanism. Applied Sciences, 2020, 10(17): 5841

[165] Wiebe J, Wilson T, Cardie C. Annotating expressions of opinions and emotions in language. Language Resources and Evaluation, 2005, 39(2): 165-210

[166] González-Carvajal S, Garrido-Merchán E C. Comparing BERT against traditional machine learning text classification. arXiv preprint arXiv:2005.13012, 2020

[167] Yang Z, Dai Z, Yang Y, et al. XLNet: Generalized autoregressive pretraining for language understanding//Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019 (NeurIPS 2019). Vancouver, Canada, 2019: 5754-5764

[168] Edwards A, Ushio A, Camacho-Collados J, et al. Guiding generative language models for data augmentation in few-shot text classification. arXiv preprint arXiv:2111.09064, 2021

[169] Sun X, Li X, Li J, et al. Text classification via large language models//Findings of the Association for Computational Linguistics(EMNLP 2023). Singapore, 2023: 8990-9005

[170] Lee J, Yoon W, Kim S, et al. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 2020, 36(4): 1234-1240

[171] Chalkidis I, Fergadiotis M, Malakasiotis P, et al. Legal-BERT: The muppets straight out of law school. arXiv preprint arXiv:2010.02559, 2020

[172] Baccianella S, Esuli A, Sebastiani F. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining//Proceedings of the International Conference on Language Resources and Evaluation (LREC 2010). Valletta, Malta, 2010

[173] Li J, Sun M. Scalable term selection for text categorization//Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning(EMNLPCoNLL 2007). Prague, Czech Republic, 2007: 774-782

[174] Khosla P, Teterwak P, Wang C, et al. Supervised contrastive learning. Advances in Neural Information Processing Systems, 2020, 33: 18661-18673

[175] Gao T, Yao X, Chen D. SimCSE: Simple contrastive learning of sentence embeddings//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021). Virtual Event/Punta Cana, Dominican Republic, 2021: 6894-6910

[176] Conneau A, Rinott R, Lample G, et al. XNLI: Evaluating cross-lingual sentence representations//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium, 2018: 2475-2485

[177] Liang P P, Zadeh A, Morency L P. Foundations and recent trends in multimodal machine learning: Principles, challenges, and open questions. arXiv preprint arXiv: 2209.03430, 2022

[178] Han X, Zhao W, Ding N, et al. PTR: Prompt tuning with rules for text classification. AI Open, 2022, 3: 182-192

[179] Min S, Lewis M, Hajishirzi H, et al. Noisy channel language model prompting for few-shot text classification//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics(ACL 2022). Volume 1: Long Papers. Dublin, Ireland, 2022: 5316-5330

[180] Petroni F, Rocktäschel T, Riedel S, et al. Language models as knowledge bases?//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019). Hong Kong, China, 2019: 2463-2473

[181] Zhang Tianjun, et al. Tempera: Test-time prompting via reinforcement learning. arXiv preprint arXiv:2211.11890, 2022

[182] Arora S, et al. Ask me anything: A simple strategy for prompting language models//Proceedings of the 11th International Conference on Learning Representations (ICLR 2023). Kigali, Rwanda, 2023: 1224-1232

[183] Help Corpus. https://www.yelp.com/dataset, 2014

[184] Sarlin P, Detone D, Malisiewicz T, et al. SuperGlue: Learning feature matching with graph neural networks//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2020). Seattle, USA, 2020: 4937-4946

[185] Guo H, Tan B, Liu Z, et al. Efficient (soft) Q-learning for text generation with limited good data//Findings of the Association for Computational Linguistics: EMNLP 2022. Abu Dhabi, Emirates, 2022: 6969-6991

[186] Li C, Liu X, Wang Y, et al. Dialogue for prompting: A policy-gradient-based discrete prompt optimization for few-shot learning. arXiv preprint arXiv:2308.07272, 2023

[187] Sutton R S, Mcallester D, Singh S, et al. Policy gradient methods for reinforcement learning with function approximation//Advances in Neural Information Processing Systems 12, NIPS Conference. Denver, USA, 1999: 1057-1063

[188] OPENAI, Achiam J, et al. GPT-4 technical report. arXiv preprint arXiv:2303.08774. DOI: 10.48550/arXiv.2303.08774, 2023

[189] Ben-David E, Oved N, Reichart R. PADA: Example-based prompt learning for on-the-fly adaptation to unseen domains. Transactions of the Association for Computational Linguistics, 2022, 10: 414-433

[190] Zubiaga A, Liakata M, Procter R, et al. Analysing how people orient to and spread rumours in social media by looking at conversational threads. PLoS One, 2016, 11(3): e0150989

[191] Williams A, Nangia N, Bowman S R. A broad-coverage challenge corpus for sentence understanding through inference//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2018). Volume 1(Long Papers). New Orleans, USA, 2018: 1112-1122

[192] Hu S, Ding N, Wang H, et al. Knowledgeable prompt-tuning: Incorporating knowledge into prompt verbalizer for text classification//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics(ACL 2022). Volume 1: Long Papers. Dublin, Ireland, 2022: 2225-2240

[193] MR Corpus. http://www.cs.cornell.edu/people/pabo/movie-reviewdata/, 2002

[194] MPQA Corpus. http://www.cs.pitt.edu/mpqa/, 2005

[195] IMDB Corpus. http://ai.stanford.edu/~amaas/data/sentiment/, 2011

[196] Amazon Corpus. https://www.kaggle.com/datasets/datafiniti/consumer-reviews-of-amazon-products, 2015

[197] Dianping Corpus. http://yongfeng.me/dataset/, 2013

[198] BDCI2019 Corpus. https://www.datafountain.cn/competitions/350/datasets, 2019

[199] NLPCC2014-SC Corpus. http://tcci.ccf.org.cn/conference/2014/pages/page04_dg.html, 2014

[200] Pontiki M, Galanis D, Papageorgiou H, et al. SemEval-2016 Task 5: Aspect based sentiment analysis//Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval@NAACLHLT 2016). San Diego, USA, 2016: 19-30

[201] Weibosenti100k Corpus. https://ieee-dataport.org/documents/weibosenti100k-and-thucnews, 2022

[202] THUCNews Corpus. http://thuctc.thunlp.org/message, 2016

[203] Sougou Corpus. https://www.sogou.com/labs/resource/tce.php, 2012
[204] ChinaNews Corpus. https://github.com/zhangxiangxiao/glyph, 2016
[205] TNEWS Corpus. https://github.com/aceimnorstuvwxz/toutiao-text-classfication-dataset, 2018
[206] Liu B, Niu D, Wei H, et al. Matching article pairs with graphical decomposition and convolutions//Proceedings of the 57th Conference of the Association for Computational Linguistics (ACL 2019). Volume 1: Long Papers. Florence, Italy, 2019: 6284-6294
[207] IFLYTEK Corpus. http://challenge.xfyun.cn/2019/gamedetail?type=detail/classifyApp, 2019
[208] Liu Z, Zhang L, Tu C, et al. Statistical and semantic analysis of rumors in Chinese social media. Scientia Sinica Informationis, 2015, 45(12): 1536
[209] LIAR Corpus. https://github.com/tfs4/liar_dataset, 2017
[210] BuzzFeedNews Corpus. https://github.com/BuzzFeedNews/2016-10-facebook-fact-check/tree/master/data, 2017
[211] CREDBANK Corpus. http://compsocial.github.io/CREDBANK-data/, 2015
[212] FacebookHoax Corpus. https://github.com/gabll/some-like-it-hoax, 2017
[213] FakeNewsNet Corpus. https://github.com/KaiDMML/FakeNewsNet/, 2019
[214] TWD Corpus. https://github.com/majingCUHK/Rumor_RvNN, 2017
[215] Fakenewsdata1 Corpus. https://github.com/rpitrust/fakenewsdata1, 2017
[216] SQuAD Corpus. https://rajpurkar.github.io/SQuAD-explorer/, 2016
[217] WikiQA Corpus. https://www.microsoft.com/en-us/download/details.aspx?id=52419, 2015
[218] Quora Corpus. https://www.kaggle.com/c/quora-question-pairs, 2017
[219] CR Corpus. https://github.com/CS287/HW1/tree/master/data, 2004
[220] cMedQA Corpus. https://github.com/zhangsheng93/cMedQA, 2017
[221] webMedQA Corpus. https://github.com/hejunqing/webMedQA, 2019
[222] XQA Corpus. https://github.com/thunlp/XQA, 2019
[223] SNLI Corpus. https://nlp.stanford.edu/projects/snli/, 2008
[224] MNLI Corpus. https://cims.nyu.edu/~sbowman/multinli/, 2018
[225] SICK Corpus. https://marcobaroni.org/composes/sick.html, 2014
[226] MSRP Corpus. https://www.microsoft.com/en-us/download/details.aspx?id=52398, 2005
[227] SciTail Corpus. https://allenai.org/data/scitail, 2017
[228] OCNLI Corpus. https://github.com/CLUEbenchmark/OCNLI, 2020
[229] Obsumed Corpus. http://davis.wpi.edu/xmdv/datasets/ohsumed.html, 2015
[230] EUR-Lex Corpus. http://www.ke.tu-darmstadt.de/resources/eurlex/eurlex.html, 2019
[231] WOS Corpus. https://data.mendeley.com/datasets/9rw3vkcfy4/2, 2017
[232] Fudan Corpus. https://www.heywhale.com/mw/dataset/5d3a9c86cf76a600360edd04, 2015
[233] Yang P, Sun X, Li W, et al. SGM: Sequence generation model for multi-label classification//Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018). Santa Fe, USA, 2018: 3915-3926
[234] Johnson A E, Pollard T J, Shen L, et al. MIMIC-III, a freely accessible critical care database. Scientific Data, 2016, 3(1): 1-9
[235] Pestian J P, Brew C, Matykiewicz P, et al. A shared task involving multi-label classification of clinical free text//Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing (BioNLP@ACL 2007). Prague, Czech Republic, 2007: 97-104
[236] Yu Y, Li M, Liu L, et al. Automatic ICD code assignment of Chinese clinical notes based on multilayer attention BiRNN. Journal of Biomedical Informatics, 2019, 91: 103114
[237] Prabowo R, Thelwall M. Sentiment analysis: A combined approach. Journal of Informetrics, 2009, 3(2): 143-157
[238] Tang D, Wei F, Qin B, et al. Sentiment embeddings with applications to sentiment analysis. IEEE Transactions on Knowledge and Data Engineering, 2015, 28(2): 496-509
[239] Cambria E. Affective computing and sentiment analysis. IEEE Intelligent Systems, 2016, 31(2): 102-107
[240] Wang Y, Sun A, Han J, et al. Sentiment analysis by capsules//Proceedings of the 2018 World Wide Web Conference (WWW 2018). Lyon, France, 2018: 1165-1174
[241] Pang B, Lee L. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales//Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005). Ann Arbor, USA, 2005: 115-124
[242] Pang B, Lee L. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts//Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics. Barcelona, Spain, 2004: 271-278
[243] Hu M, Liu B. Mining and summarizing customer reviews//Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Seattle, USA, 2004: 168-177
[244] Thelwall M, Buckley K, Paltoglou G. Sentiment strength detection for the social web. Journal of the American Society for Information Science and Technology, 2012, 63(1): 163-173
[245] Nakov P, Ritter A, Rosenthal S, et al. SemEval-2016 task 4: Sentiment analysis in Twitter. arXiv preprint arXiv: 1912.01973, 2019

[246] Meng Y, Wu W, Wang F, et al. Glyce: Glyph-vectors for Chinese character representations//Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019 (NeurIPS 2019). Vancouver, Canada, 2019: 2742-2753

[247] Quan C, Ren F. A blog emotion corpus for emotional expression analysis in Chinese. Computer Speech & Language, 2010, 24(4): 726-749

[248] Barberà P, Boydstun A E, Linn S, et al. Automated text classification of news articles: A practical guide. Political Analysis, 2021, 29(1): 19-42

[249] Li C, Zhan G, Li Z. News text classification based on improved Bi-LSTM-CNN//Proceedings of the 2018 9th International Conference on Information Technology in Medicine and Education (ITME). Hangzhou, China, 2018: 890-893

[250] Sun C, Qiu X, Xu Y, et al. How to fine-tune BERT for text classification?//Proceedings of the 18th China National Conference on Chinese Computational Linguistics (CCL 2019). Kunming, China, 2019: 194-206

[251] Shu K, Sliva A, Wang S, et al. Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorations Newsletter, 2017, 19(1): 22-36

[252] Pérez-Rosas V, Kleinberg B, Lefevre A, et al. Automatic detection of fake news//Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018). Santa Fe, USA, 2018: 3391-3401

[253] Reis J C, Correia A, Murai F, et al. Supervised learning for fake news detection. IEEE Intelligent Systems, 2019, 34(2): 76-81

[254] Shu K, Cui L, Wang S, et al. dEFEND: Explainable fake news detection//Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD 2019). Anchorage, USA, 2019: 395-405

[255] Song C, Yang C, Chen H, et al. CED: Credible early detection of social media rumors. IEEE Transactions on Knowledge and Data Engineering, 2021, 33(8): 3035-3047

[256] Shu K, Mahudeswaran D, Wang S, et al. FakeNewsNet: A data repository with news content, social context and dynamic information for studying fake news on social media. arXiv preprint arXiv:1809.01286, 2018

[257] Shu K, Wang S, Liu H. Exploiting tri-relationship for fake news detection. arXiv preprint arXiv:1712.07709, 2017

[258] Ma J, Gao W, Wong K. Rumor detection on Twitter with tree-structured recursive neural networks//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018). Volume 1: Long Papers. Melbourne, Australia, 2018: 1980-1989

[259] Horne B, Adali S. This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. arXiv preprint arXiv:1703.09398, 2017

[260] Rajpurkar P, Zhang J, Lopyrev K, et al. SQuAD: 100,000+ questions for machine comprehension of text//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016). Austin, USA, 2016: 2383-2392

[261] Yang Y, Yih W, Meek C. WikiQA: A challenge dataset for open-domain question answering//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015). Lisbon, Portugal, 2015: 2013-2018

[262] Zhang S, Zhang X, Wang H, et al. Chinese medical question answer matching using end-to-end character-level multi-scale CNNs. Applied Sciences, 2017, 7(8): 767

[263] Zhang S, Zhang X, Wang H, et al. Multi-scale attentive interaction networks for Chinese medical question answer selection. IEEE Access, 2018, 6: 74061-74071

[264] Liu J, Lin Y, Liu Z, et al. XQA: A cross-lingual open-domain question answering dataset//Proceedings of the 57th Conference of the Association for Computational Linguistics (ACL 2019). Volume 1: Long Papers. Florence, Italy, 2019: 2358-2368

[265] Wang S, Jiang J. Learning natural language inference with LSTM//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, USA, 2016: 1442-1451

[266] Gong Y, Luo H, Zhang J. Natural language inference over interaction space//Proceedings of the 6th International Conference on Learning Representations (ICLR 2018). Conference Track Proceedings. Vancouver, Canada, 2018

[267] Chen Q, Zhu X, Ling Z, et al. Enhanced LSTM for natural language inference//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017). Volume 1: Long Papers. Vancouver, Canada, 2017: 1657-1668

[268] Bentivogli L, Bernardi R, Marelli M, et al. SICK through the SemEval glasses. Lesson learned from the evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. Language Resources and Evaluation, 2016, 50: 95-124

[269] Bowman S R, Angeli G, Potts C, et al. A large annotated corpus for learning natural language inference//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015). Lisbon, Portugal, 2015: 632-642

[270] Dolan B, Quirk C, Brockett C. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources//Proceedings of the 20th International Conference on Computational Linguistics. Geneva, Switzerland, 2004

[271] Dagan I, Glickman O, Magnini B. The PASCAL recognising textual entailment challenge//Machine Learning Challenges, Evaluating Predictive Uncertainty, Visual Object Classification and Recognizing Textual Entailment, First PASCAL Machine Learning Challenges Workshop (MLCW 2005). Lecture Notes in Computer Science: Volume 3944. Southampton, UK, 2005: 177-190

[272] Khot T, Sabharwal A, Clark P. SciTail: A textual entailment dataset from science question answering//Proceedings

of the 32nd AAAI Conference on Artificial Intelligence (AAAI-18), the 30th Innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18). New Orleans, USA, 2018: 5189-5197

[273] Hu H, Richardson K, Xu L, et al. OCNLI: Original Chinese natural language inference//Findings of the Association for Computational Linguistics: EMNLP 2020. Online Event, 2020: 3512-3526

[274] Lehmann J, Isele R, Jakob M, et al. DBpedia—A large-scale, multilingual knowledge base extracted from Wikipedia. Semantic Web, 2015, 6(2): 167-195

[275] Kowsari K, Brown D E, Heidarysafa M, et al. HDLTex: Hierarchical deep learning for text classification//Proceedings of the 16th IEEE International Conference on Machine Learning and Applications (ICMLA 2017). Cancun, Mexico, 2017: 364-371

[276] Amazon670K Corpus. http://manikvarma.org/downloads/XC/XMLRepository.html, 2016

[277] Wang F, Wang Z, Li Z, et al. Concept-based short text classification and ranking//Proceedings of the 23rd ACM International Conference on Information and Knowledge Management (CIKM 2014). Shanghai, China, 2014: 1069-1078

[278] BDCI-2018. https://github.com/yilifzf/BDCI_Car_2018

[279] Xie P, Shi H, Zhang M, et al. A neural architecture for automated ICD coding//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018). Volume 1: Long Papers. Melbourne, Australia, 2018: 1066-1076

[280] Shi H, Xie P, Hu Z, et al. Towards automated ICD coding using deep learning. arXiv preprint arXiv: 1711.04075, 2017

[281] Li M, Fei Z, Zeng M, et al. Automated ICD-9 coding via a deep learning approach. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2018, 16(4): 1193-1202

[282] Kaur R, Ginige J A, Obst O. A systematic literature review of automated ICD coding and classification systems using discharge summaries. arXiv preprint arXiv: 2107.10652, 2021

[283] Lee J Y, Dernoncourt F. Sequential short-text classification with recurrent and convolutional neural networks//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, USA, 2016: 515-520

[284] Lever J. Classification evaluation: It is important to understand both what a classification metric expresses and what it hides. Nature Methods, 2016, 13(8): 603-605

[285] Yacouby R, Axman D. Probabilistic extension of precision, recall, and F1 score for more thorough evaluation of classification models//Proceedings of the 1st Workshop on Evaluation and Comparison of NLP Systems. Online, 2020: 79-91

[286] Provost F J, Fawcett T, Kohavi R. The case against accuracy estimation for comparing induction algorithms//Proceedings of the 15th International Conference on Machine Learning (ICML 1998). Madison, USA, 1998: 445-453

[287] Davis J, Goadrich M. The relationship between precision-recall and ROC curves//ACM International Conference Proceeding Series: Volume 148 Machine Learning, Proceedings of the 23rd International Conference (ICML 2006). Pittsburgh, USA, 2006: 233-240

[288] Wang G, Yang M, Zhang L, et al. Momentum accelerates the convergence of stochastic AUPRC maximization//Proceedings of the Machine Learning Research: Volume 151 International Conference on Artificial Intelligence and Statistics (AISTATS 2022). Virtual Event, 2022: 3753-3771

[289] Myerson J, Green L, Warusawitharana M. Area under the curve as a measure of discounting. Journal of the Experimental Analysis of Behavior, 2001, 76(2): 235-243

[290] Wang W, Yang N, Wei F, et al. Gated self-matching networks for reading comprehension and question answering//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017). Volume 1: Long Papers. Vancouver, Canada, 2017: 189-198

[291] Severyn A, Moschitti A. Learning to rank short text pairs with convolutional deep neural networks//Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. Santiago, Chile, 2015: 373-382

[292] Benjamini Y. Discovering the false discovery rate. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2010, 72(4): 405-416

[293] Zafar M B, Valera I, Gomez-Rodriguez M, et al. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment//Proceedings of the 26th International Conference on World Wide Web (WWW 2017). Perth, Australia, 2017: 1171-1180

[294] Mccallum A K. Multi-label text classification with a mixture model trained by EM//Proceedings of the AAAI 99 Workshop on Text Learning. Orlando, USA, 1999

[295] Schütze H, Manning C D, Raghavan P. Introduction to Information Retrieval: Volume 39. Cambridge, UK: Cambridge University Press, 2008

[296] Schapire R E, Singer Y. Improved boosting algorithms using confidence-rated predictions. Machine Learning, 1999, 37(3): 297-336

[297] Liu J, Chang W, Wu Y, et al. Deep learning for extreme multi-label text classification//Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. Shinjuku, Japan, 2017: 115-124

[298] Järvelin K, Kekäläinen J. Cumulated gain-based evaluation of IR techniques. ACM Transactions on Information Systems, 2002, 20(4): 422-446

[299] Iki T, Aizawa A. Effect of visual extensions on natural language understanding in vision-and-language models//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021). Virtual Event/Punta Cana, Dominican Republic, 2021: 2189-2196

[300] Reddy T, Williams R, Breazeal C. Text classification for AI education//Proceedings of the 52nd ACM Technical Symposium on Computer Science Education. Virtual Event, USA, 2021: 1381

[301] Liang X, Cheng D, Yang F, et al. F-HMTC: Detecting financial events for investment decisions based on neural hierarchical multi-label text classification//Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI 2020). Yokohama, Japan, 2020: 4490-4496

[302] Shanthi P, Modi S, Hareesha K, et al. Classification and comparison of malignancy detection of cervical cells based on nucleus and textural features in microscopic images of uterine cervix. International Journal of Medical Engineering and Informatics, 2021, 13(1): 1-13

[303] Zhu X, Goldberg A B. Introduction to semi-supervised learning. Synthesis Lectures on Artificial Intelligence and Machine Learning, 2009, 3(1): 1-130

[304] Wang Y, Yao Q, Kwok J T, et al. Generalizing from a few examples: A survey on few-shot learning. ACM Computing Surveys, 2020, 53(3): 1-34

[305] Hirschberg J, Manning C D. Advances in natural language processing. Science, 2015, 349(6245): 261-266

[306] Rojas K R, Bustamante G, Oncevay A, et al. Efficient strategies for hierarchical text classification: External knowledge and auxiliary tasks//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020). Online, 2020: 2252-2257

[307] Miller G A. WordNet: A lexical database for English. Communications of the ACM, 1995, 38(11): 39-41

[308] Yin W, Shen L. A short text classification approach with event detection and conceptual information//Proceedings of the 2020 5th International Conference on Machine Learning Technologies (ICMLT'20). New York, USA, 2020: 129-135

[309] Majewski P, Szymanski J. Text categorization with semantic commonsense knowledge: First results//Proceedings of the 14th International Conference on Neural Information Processing (ICONIP 2007). Lecture Notes in Computer Science: Volume 4985. Kitakyushu, Japan, 2007: 769-778

[310] Pujara J, Miao H, Getoor L, et al. Knowledge graph identification//The Semantic Web-ISWC 2013-12th International Semantic Web Conference. Lecture Notes in Computer Science: Volume 8218. Sydney, Australia, 2013: 542-557

[311] Gao J, Li P, Chen Z, et al. A survey on deep learning for multimodal data fusion. Neural Computation, 2020, 32(5): 829-864

[312] Chen J, Hu Y, Liu J, et al. Deep short text classification with knowledge powered attention//Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI 2019), the 31st Innovative Applications of Artificial Intelligence Conference (IAAI 2019), the 9th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI 2019). Honolulu, USA, 2019: 6252-6259

[313] Pappas N, Popescu-Belis A. Multilingual hierarchical attention networks for document classification//Proceedings of the 8th International Joint Conference on Natural Language Processing (IJCNLP 2017). Volume 1: Long Papers. Taipei, China, 2017: 1015-1025

[314] Chang W, Yu H, Zhong K, et al. Taming pretrained transformers for extreme multi-label text classification//Proceedings of the 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'20). Virtual Event, USA, 2020: 3163-3171

[315] Wu T, Huang Q, Liu Z, et al. Distribution-balanced loss for multilabel classification in long-tailed datasets//Proceedings of the 16th European Conference on Computer Vision (ECCV 2020). Lecture Notes in Computer Science: Volume 12349. Glasgow, UK, 2020: 162-178

[316] Gong J, Teng Z, Teng Q, et al. Hierarchical graph transformer based deep learning model for large-scale multi-label text classification. IEEE Access, 2020, 8: 30885-30896

[317] Menghani G. Efficient deep learning: A survey on making deep learning models smaller, faster, and better. ACM Computing Surveys, 2023, 55(12): 259:1-259:37

[318] Akpatsa S K, Li X, Lei H. A survey and future perspectives of hybrid deep learning models for text classification//Proceedings of the 7th International Conference on Artificial Intelligence and Security (ICAIS 2021). Lecture Notes in Computer Science: Volume 12736. Dublin, Ireland, 2021: 358-369

[319] Ji Shou-Ling, Du Tian-Yu, Li Jin-Feng, et al. Security and privacy of machine learning models: A survey. Journal of Software, 2021, 32(1): 41-67(in Chinese)
(纪守领, 杜天宇, 李进锋等. 机器学习模型安全与隐私研究综述. 软件学报, 2021, 32(1): 41-67)

[320] Liu Xiao-Ming, Zhang Zhao-Han, Yang Chen-Yang, et al. Adversarial technology of text content on online social networks. Chinese Journal of Computers, 2022, 45(8): 1571-1597(in Chinese)
(刘晓明, 张兆晗, 杨晨阳等. 在线社交网络文本内容对抗技术. 计算机学报, 2022, 45(8): 1571-1597)

[321] Xu J, Chen J, You S, et al. Robustness of deep learning models on graphs: A survey. AI Open, 2021, 2: 69-78

[322] Chakraborty S, Tomsett R, Raghavendra R, et al. Interpretability of deep learning models: A survey of results//Proceedings of the 2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI 2017). San Francisco, USA, 2017: 1-6

[323] Peng Baolin, et al. Instruction tuning with GPT-4. arXiv preprint arXiv:2304.03277, 2023

[324] Kaplan J, Mccandlish S, Henighan T, et al. Scaling laws for neural language models. arXiv preprint arXiv: 2001.08361, 2020

[325] Wiegreffe S, Pinter Y. Attention is not not explanation// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019). Hong Kong, China, 2019: 11-20

[326] Sap M, Card D, Gabriel S, et al. The risk of racial bias in hate speech detection//Proceedings of the 57th Conference of the Association for Computational Linguistics (ACL 2019). Volume 1: Long Papers. Florence, Italy, 2019: 1668-1678

[327] Bolukbasi T, Chang K, Zou J Y, et al. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings//Proceedings of the Annual Conference on Neural Information Processing Systems 2016. Barcelona, Spain, 2016: 4349-4357

LIU Xiao-Ming, Ph. D., associate professor. His research interests include online social network adversarial technology, spatio-temporal anomaly detection, machine learning and its applications.

LI Cheng-Zheng-Xu, M. S. candidate. His research interests include reinforcement learning, text intelligent computing.

WU Shao-Cong, M. S. candidate. His research interests include recommendation systems, knowledge graph mining, text intelligent computing.

ZHANG Yu-Chen, Ph. D. candidate. His research interests include online social network adversarial technology and fake news detection.

BAI Hong-Yan, M. S. candidate. Her research interests include community detection, graph neural network applications, text intelligent computing.

CHENG Ze-Hua, M. S. candidate. Her research interests include machine generated text detection, text intelligent computing.

CHEN Zhuo, M. S. candidate. His research interests include graph neural network applications, text intelligent computing.

LI Yong-Feng, Ph. D. candidate. His research interests include anomaly detection.

LAN Yu, Ph. D., assistant professor. Her research interests include system optimization, machine learning and its applications.

SHEN Chao, Ph. D., professor. His research interests are trusted artificial intelligence and artificial intelligence security, cyber-physical system control and security, software security and testing.

Background

In recent years, with the advent of the era of big data, the text information in the Internet has ushered in a blowout growth. As one of the most important technologies in natural language processing, text classification has a wide range of applications, and related research methods and ideas are constantly evolving. This paper conducts an overall survey of traditional machine learning methods and emerging deep learning methods, and comprehensively discusses the technical challenges faced by current methods and future research directions. Specifically, this paper consists of seven parts, which are: (1) Introducing the relevant basic knowledge of text classification technology. (2) Summarizing the text classification methods based on traditional machine learning with discussions of their advantages and disadvantages. (3) Introducing the text classification methods based on deep learning, which are classified according to the adopted different basic network architectures. (4) In order to facilitate readers to verify the validity of the model, this paper systematically summarizes the relevant datasets for the seven most widely used scenarios of text classification technology. (5) this paper introduces the commonly used model evaluation methods under different task objectives in detail, so as to quantitatively evaluate the model performance. (6) Based on the above, this paper summarizes and compares the performance of different types of text classification algorithms in typical application scenarios. (7) Summarizing the challenges faced by text classification technology and the important research directions in the future from two aspects, e. g., data limitation, model computation performance. This paper can effectively address the gaps of the text classification method survey in unbalanced data, small sample data and other application scenarios, and provide a reference for the research on related issues.

This work is supported by the National Key Research and Development Program (No. 2020YFB1406900), the National Natural Science Foundation of China (Nos. 62272371, 61902308, U21B2018, 62103323, 62161160337, 61822309, 61773310), the Initiative Postdocs Supporting Program (Nos. BX20190275, BX20200270), the China Postdoctoral Science Foundation (Nos. 2019M663723, 2021M692565), the Fundamental Research Funds for the Central Universities under Grant (No. xzy012024144) and the Shaanxi Province Key Industry Innovation Program (No. 2021ZDLGY01-02).