计算机科学
COMPUTER SCIENCE

融合主题特征的文本情感分析模型

杨俊哲, 宋莹, 陈逸菲

引用本文

杨俊哲, 宋莹, 陈逸菲. 融合主题特征的文本情感分析模型[J]. 计算机科学, 2024, 51(6A): 230600111-8.

YANG Junzhe, SONG Ying, CHEN Yifei. Text Emotional Analysis Model Fusing Theme Characteristics [J]. Computer Science, 2024, 51(6A): 230600111-8.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230600111

融合主题特征的文本情感分析模型

杨俊哲[1]　宋　莹[2]　陈逸菲[2]
1 南京信息工程大学自动化学院　南京 210044
2 无锡学院自动化学院　江苏 无锡 214105
(20211249590@nuist.edu.cn)

摘　要　随着大型语言模型的快速发展,如何在保证模型性能的同时减少模型参数量,成为了自然语言处理领的一个重要挑战。然而,现有的参数压缩技术往往难以兼顾模型的稳定性和泛化能力。为此,提出了一种融合主题特征的情感分析新架构,旨在利用主题信息增强模型对文本情感极性的判断能力。具体而言,采用一种结合 LDA 和 K-means 的方法来提取文本的主题特征,并将其作为固定维度的向量与词嵌入进行拼接,得到新的词向量表示。随后使用平均池化技术构建句子级别的表征向量,并输入到一个全连接层进行情感分类。为了验证所提模型的有效性,在公开的情感分析数据集上与多个基准算法进行了对比实验。实验结果表明,所提模型在多个数据集上明显优于 ALBERT,准确率提高了约 3.5%,在参数量仅有微小增加的情况下维持了较高的稳定性和泛化能力。

关键词:情感分析;ALBERT 模型;LDA 模型;主题特征;平均池化
中图分类号　TP391

Text Emotional Analysis Model Fusing Theme Characteristics

YANG Junzhe[1],SONG Ying[2] and CHEN Yifei[2]
1 School of Automation,Nanjing University of Information Science and Technology,Nanjing 210044,China
2 School of Automation,Wuxi University,Wuxi,Jiangsu 214105,China

Abstract　With the rapid development of large-scale language models,how to reduce the number of model parameters while ensuring model performance has become an important challenge in the field of natural language processing. However,the existing parameter compression techniques are often difficult to balance the stability and generalization ability of the model. To this end,this paper proposes a new framework for sentiment analysis that integrates topic features,aiming to use topic information to enhance the model's ability to judge text sentiment polarity. Specifically,a method combining LDA and K-means is used to extract the topic features of the text,and it is spliced with word embeddings as a fixed-dimensional vector to obtain a new word vector representation. Sentence-level representation vectors are then constructed using average pooling techniques and fed into a fully connected layer for sentiment classification. To verify the effectiveness of the proposed model,comparative experiments with multiple benchmark algorithms are carried out on public sentiment analysis datasets. Experimental results show that the proposed model is significantly better than ALBERT in multiple data sets,with an accuracy rate increases by about 3.5%,and it maintains high stability and generalization ability with only a small increase in the number of parameters.

Keywords　Emotional analysis,ALBERT model,Latent dirichlet allocation,Theme features,Average pooling

1　引言

情感分析是自然语言处理领域的一个关键任务[1],旨在收集和分析人们对各种主题、产品、事件和服务表达的意见,在社交媒体、电子商务[2]和在线教育[3]等诸多领域有着广泛应用。随着网络社交媒体的发展,用户通过网络表达观点形成舆情已成为一种常态,进而产生了大量数据[4]。公司、政府等机构可以通过分析数据挖掘人们的意见和情感倾向,对市场决策和舆情分析具有重要价值[5]。因此,近几年情绪分析不仅在研究人员中,而且在企业、政府和一些组织中也得到了广泛认可。

然而,情感分析和评估过程也存在着许多挑战,这些挑战影响了情绪的准确解读和极性判断[6]。具体而言,在社交媒体领域,文本数据通常是由非结构化、含有大量网络用语和隐含情感(如隐喻、反讽等)的散句组成,这些特征使得文本表示阶段难以提取有效的特征信息,从而影响了向量表示和情感分类的准确性和效率。

为了提高词向量的表征能力,近年来出现了一些基于深度学习的方法,并出现了大量预训练模型[7]。如 BERT 利用 Transformer 编码器结构学习了上下文相关的词向量,显著提

基金项目:江苏省高等学校自然科学研究面上项目(19KJB520044);江苏省研究生实践创新计划项目(SJCX23_0392)
This work was supported by the Natural Science Foundation of the Jiangsu Higher Education Institutions of China(19KJB520044) and Postgraduate Research & Practice Innovation Program of Jiangsu Province(SJCX23_0392).
通信作者:宋莹(sypeace@126.com)

升了自然语言处理任务的性能[8]。但 BERT 巨大的参数量也带来了训练、推理和存储的成本上升。ALBERT 采用了内部参数共享的策略，使得其参数量相比 BERT 减少了 90%[9]。然而，参数共享策略的不稳定性导致模型表征能力下降，从而影响了情感分析任务的准确性和效率，而这些任务又高度依赖于词向量对文本语义的刻画。

因此，本文探索了如何利用文本主题信息来提高 ALBERT 在情感分析任务上的稳定性和性能，使模型在维持较少参数量的前提下高提准确率。本文提出了一种基于主题特征和 ALBERT 的情感分析模型新架构，并在多个数据集上进行了实验验证。具体来说，首先采用一种结合 LDA 模型和 K-means 算法的方法来提取文本的主题特征，并将其作为固定维度的向量与词嵌入进行拼接，从而丰富词向量对文本语义的刻画；然后使用表征向量平均池化的技术来构建句子级别的表征向量，从而提高情感分类任务的准确率；最后在多个数据集上与若干基准算法进行对比实验，证明了本文提出的模型在文本情感分析领域具有优异的效果。

2 相关工作

情感分析的常用方法主要分为基于词典、基于机器学习和基于深度学习的方法。

2.1 基于情感词典

基于词典的方法主要利用预先构建好的情感词典，情感词典是由一些带有正面或负面倾向的词语组成的集合。该方法实现简单且不依赖于标注数据，只需将经过预处理（如分词、去停用词等）后的文本同预设的情感词典进行匹配，然后根据匹配结果计算文本的情感得分并判断极性[10]。然而，基于词典的方法也存在一些局限性，如难以适应新出现或多义的情感词语，难以捕捉上下文和语境对情感倾向的影响，难以处理否定、反讽等复杂语言现象。因此，基于词典的方法需要不断地更新和维护情感词典，增加人工成本和先验知识。

2.2 基于机器学习

基于机器学习的方法主要是利用一些有监督或无监督的算法来从标注或未标注数据中学习情感分类模型，常用的算法有支持向量机（SVM）、朴素贝叶斯（NB）、随机森林（RF）、最大熵（ME）等[11]。基于机器学习的方法相比基于词典的方法，可以更好地适应不同领域和语言的情感分析任务，提高分类准确率和鲁棒性。然而，基于机器学习的方法也面临一些挑战和问题，如模型训练需要大量高质量的标注数据，而标注数据往往难以获取和保证一致性；模型训练需要合适的特征提取和选择方法，而特征提取和选择往往依赖于人工经验和先验知识；模型训练需要合理的参数调整和优化方法，而参数调整和优化往往需要耗费大量时间和计算资源。

2.3 基于深度学习

基于深度学习的方法主要是利用一些神经网络模型来从大规模的未标注数据中自动学习情感分类模型。常用的模型有卷积神经网络（CNN）[12]、循环神经网络（RNN）[13]、长短期记忆网络（LSTM）[14]、注意力机制（Attention）[15]等。基于深度学习的方法相比基于机器学习的方法，可以更好地捕捉文本中的语义和情感信息，提高分类准确率和泛化能力[16]。然而，基于深度学习的方法也面临一些挑战和问题，如模型训练需要大量的计算资源和时间，模型结构和参数难以解释和理解，模型性能难以评估和优化等。此外，基于深度学习的方法还需要解决一些特殊的语言现象，如一词多义、反讽、否定等，这些现象会影响情感倾向的判断。

Google 开放了词嵌入模型 Word2vec[17]，利用神经网络对大规模语料库进行训练，得到每个词的低维稠密向量表示，但是 Word2vec 无法解决一词多义以及语言歧义等问题。

Vaswani 等[18]提出了 Transformer 结构，利用自注意力机制取代传统的循环神经网络或卷积神经网络，有效地捕捉了序列中不同位置之间的依赖关系，缓解了困扰自然语言处理领域多年的长距离依赖问题。Huang 等[19]在情感分析任务中引入了注意力机制，有效提升了模型对隐式情感的敏感度。Google 提出了 BERT 模型[8]，该模型是第一个基于无监督预训练的深度双向语言表示模型，使用了掩码语言模型和下一个句子预测作为预训练任务，使得 BERT 在 GLUE 基准测试中 11 项 NLP 任务上取得了最佳性能。BERT 模型利用下游任务来捕捉句子之间的语义关系，从而提高了语言表示模型的泛化能力[20]。其优势在于微调简单，只需要在预训练好的 BERT 模型上添加一个额外的输出层即可适应不同的下游 NLP 任务。例如，针对情感分析任务中，只需在预训练好的 BERT 模型上添加一个分类器层，就可以获得具体任务的语言表示。Li 等[21]将 BERT 模型应用于股票投资者的情感分析任务，结果表明其准确率高于 LSTM 和 SVM。由于 BERT 模型具有很强的泛化性，其参数量以及预训练时间也相应地很大，最小的 BERT 预训练模型参数量也有一亿多。Song 等[22]首次将 BERT 运用在微博文本情感分析领域，并通过实验分析证明了 BERT 更容易学习边界信息，但是 BERT 模型庞大的参数量仍是个问题。为减少模型参数，提高模型精确度，Lan 等[9]提出了 ALBERT 模型，在 BERT 的基础上使用了两种参数压缩技术：分解嵌入参数化和跨层参数共享，显著降低了预训练对内存和计算资源的消耗，而且训练时长缩短了 20%。研究发现，在小样本情况下，使用预训练好的 ALBERT 微调可以有效地利用其学习到的通用语言特征，提高文本情感分类的效果。Wang 等[23]提出了 ALBERT-LSTM 模型，该模型在 ALBERT 的基础上增加了一个 LSTM 层，用于捕捉序列中远距离的语义特征，从而提升情感分析的效果。然而，ALBERT 本身就具有强大的语义表示能力，可以通过足够深的 Transformer 网络以及自注意力机制和位置嵌入来解决“一词多义”和长距离依赖问题，所以该模型增加的 LSTM 层的必要性和有效性有待验证。Gao 等[24]提出 ALBERT-TextCNN-Attention 模型，弥补了轻量级 ALBERT 模型与 BERT 模型相比精度略有损失的问题，但固定大小的卷积核使得一些较长或较短的文本可能无法被有效地处理。

为了适应不同领域的数据分布和特点，许多研究者采用了领域适应（domain adaptation）或领域迁移（domain transfer）的方法，在通用预训练模型（如 BERT）的基础上，对目标任务进行进一步预训练，以增强模型对目标领域数据的理解和适应能力。大量实验证明，这种方法可以显著提升模型在目标领域任务上的效果。

2.4 主题提取模型

David 等[25]提出了一种基于潜在迪利克雷分布（LDA）的生成概率模型，该模型可以从文档中自动发现主题，并具有良

好的模块化和可扩展性，可以与其他复杂的模型结合使用。在该模型中，每篇文档都可以看作是多个不同主题的混合，每个主题都可以用有限词汇表中词语的概率分布来描述。Yan等[26]提出了一种结合LDA和情感词典的情感分类模型，能够有效地建立词、句和文档主题之间的关联。Xue等[27]使用LDA模型对Twitter数据集进行情感极性分析，模型使用加权统计法TF-IDF生成的词向量作为输入特征，但由于模型的局限性，无法捕捉上下文语序与情感极性的联系，从而导致模型泛化能力不足。K-means是一种经典的划分式聚类算法，适用于高维度数据集的聚类，该算法能够将向量空间中的数据点划分为多个相似的簇。Biu等[28]提出了一种结合LDA和K-means的方法，通过实验证明了其在聚类性能上的优势。文档数据集可用高维矩阵表示，其中每个向量代表一个文档。因此，可以使用向量之间的距离来表示两个文档之间的情感相似或差异度。由于K-means对高维数据的聚类效果不佳，通常需要降低文档维度并使用合适的距离度量来提升聚类效果。在LDA中，每篇文档都可以看作是多个不同主题的混合，每个主题都由有限词汇表中的概率分布表征来描述。然而，LDA也存在一些局限性，比如无法解决“一词多义”问题，并且无法充分考虑字/词顺序对语义情感的影响。

3 融合主题信息的情感分析模型

本文提出了一种结合ALBERT、LDA和K-means的LK-ALBERT模型，用于对中文文本的情感极性分类。结构如图4所示，该模型主要包括预处理层、嵌入层、编码层和分类层4个部分。

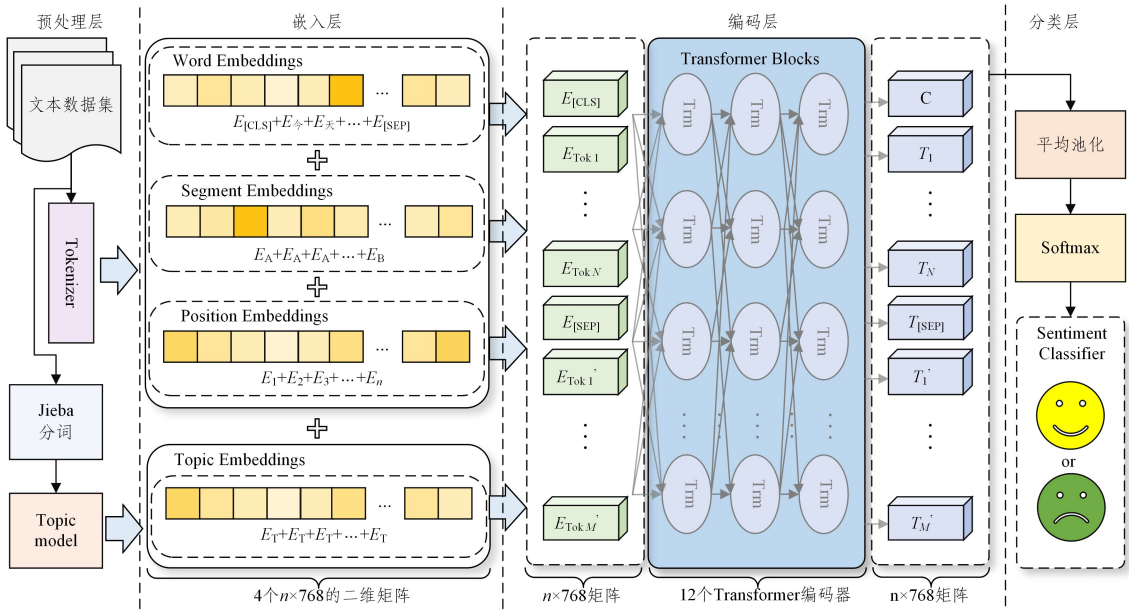


图1 模型结构

Fig. 1 Model structure

3.1 预处理层

预处理层主要包括文本清洗、文本分词和主题特征提取3个步骤。

3.1.1 文本清洗

文本清洗是指去除文本中的无关信息，使其更适合后续的分析，其主要步骤如下：

1）规范化（删除数字、标点符号、乱码、URL链接以及停用词等）；

2）将表情符号替换为对应文本形式，将繁体字转换为简体；

3）将序列填充或截断为固定长度。

3.1.2 文本分词

文本分词是将预处理后的文档D经过嵌入层的SentencePiece分词器转化为长度为n的token序列。在每个文本的开头添加特殊token“[CLS]”，在每个句子末尾添加特殊token“[SEP]”，并使用“[PAD]”填充序列到固定的长度。这些特殊token有助于编码层了解文本的结构和上下文特征。

3.1.3 主题特征提取

主题特征提取是利用LDA结合K-means从文本中抽取主题信息并将其表示为$n\times768$的主题向量。

该方法可以有效地利用LDA模型的概率分布特性，挖掘出文本的隐含主题信息，并将其转化为向量表示，从而增强文本特征的表达能力。同时实现主题信息的自适应抽取，并将其与编码层其他特征进行拼接，从而增强文本特征的融合能力。

主题特征的提取任务依赖LDA和K-means的融合算法。其中，LDA是以概率分布的形式给数据集中每个文档分配主题的三层贝叶斯统计学模型，能够挖掘出文本的隐含主题信息。其原理是假设每篇文档由多个主题组成，每个主题由多个词语组成，模型通过迭代更新每个主题下的词语分布和每篇文档下的主题分布，最终得到每个词语属于不同主题的概率和每篇文档包含不同主题的概率。因此，为使LDA输出文档的主题向量，本文针对LDA做了以下改进：

1）将聚类用的LDA拓展为能够生成文档向量的主题模型，不再使用LDA模型的聚类结果；

2）将“文档-主题”的Dirichlet分布修改为以主题作为维度，以概率值作为维度上的数值，从而得到文档在不同主题维度上的向量表示；

3）结合LDA主题模型和K-means算法，增强模型对主题的抽取能力；

4）将每篇文档的主题向量拼接成一个大小为$n\times768$的矩阵。

改进后的 LK 模型最外层是文档集合层，然后是文档层和词层。其中，单个圆圈代表潜在变量，矩形表示重复采样（右下角字母为重复采样的次数），箭头表示两个变量之间条件概率的依赖关系。文档的主题分布和主题的词语分布分别为 $\boldsymbol{\theta}$ 和 $\boldsymbol{\varphi}$，它们都服从先验 Dirichlet 分布，其语料级参数为 $\boldsymbol{\alpha}$ 和 $\boldsymbol{\beta}$。文档集合为 $D=\{d_i \mid i\in\{1,2,\cdots,M\}\}$，每篇文档中的句子为 d，句子的编号为 i，每个句子中汉字数量为 N，汉字编号为 n。每篇文档分配给词的隐含主题份额为 z，文本的主题词为 w，每一个词都有一个潜在主题，主题个数为 K，其文档的主题分布如式(1)所示：

$$P(\boldsymbol{w},\boldsymbol{z},\boldsymbol{\theta}\mid\boldsymbol{\alpha},\boldsymbol{\beta})=P(\boldsymbol{\theta}\mid\boldsymbol{\alpha})\prod_{n=1}^{N}P(\boldsymbol{z}_n\mid\boldsymbol{\theta})P(\boldsymbol{w}_n\mid\boldsymbol{z}_n,\boldsymbol{\beta}) \tag{1}$$

K-means 聚类算法的基本逻辑是先定义质心个数 C，然后对象经过不断循环分配到最近的质心。在每一步中，都需要重新计算 C 个新的质心，然后重新分配对象，直到不再进行任何更改。在本文改进的算法中，质心即为主题，所以 $C=K$。为了提高 LDA 提取主题信息的准确率，本文在 LDA 算法中融合 K-means 聚类算法，将 $\boldsymbol{\theta}$ 作为 K-means 聚类算法的输入。LK 模型的具体结构如图 2 所示。

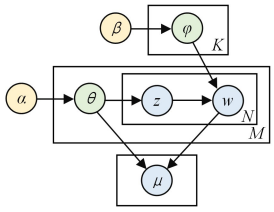


图 2 LDA-K-means 结构

Fig. 2 LDA-K-means structure

3.2 嵌入层

通过词嵌入模型获得高质量词向量是利用深度学习对文本的情感极性进行分类的前提。为了对评论数据集的情感进行分类，首先需要将其转化为向量。嵌入层的作用，是将 token 序列映射为 4 个固定维度的向量表示，实现文本向量化。

3.2.1 词嵌入

$\boldsymbol{\omega}$ 为 $n\times768$ 的词嵌入（word embeddings）是将每个汉字 token 映射到一个高维空间中，转化为固定向量表示，以便后续编码层的处理任务。使用字片段（wordpiece）作为基本单元，能够有效解决一词多义和陌生词的问题，使得向量可以在保持相对较低的维度下，尽可能多地包含复杂语义关系。

3.2.2 段嵌入

$\boldsymbol{\delta}$ 为 $n\times768$ 的段嵌入（segment embeddings），是文章段落的向量表示。因为文档中句子的顺序对于情感的表达至关重要，所以添加句段位置来辅助模型区分两个不同的句子顺序或段落顺序。该部分是一个二值向量，一篇文档 D_i 的段嵌入表示如式(2)所示：

$$\begin{cases}\boldsymbol{\delta}(seg,2i+1)=[E_a,E_b,E_a,\cdots,E_a]\\ \boldsymbol{\delta}(seg,2i)=[E_a,E_b,E_a,\cdots,E_b]\end{cases} \tag{2}$$

3.2.3 位置嵌入

如果仅使用注意力机制是无法捕获序列的顺序信息的，这是该机制的一个短板，也就是顺序免疫。因此，引入 $\boldsymbol{\rho}$ 为 $n\times768$ 的位置嵌入（position embeddings），用于表示输入序列中每个词的位置信息，引导编码层来学习文本序列的位置特征，弥补注意力机制的短板。与词嵌入类似，字的位置信息被映射到序列空间中的一个点，然后以数值向量的形式表示出来。位置嵌入使得编码层在维持对文本并行处理的同时，学习文本的时间序列信息。但如果给每个时间步分配一个从 0 到 n 的数字来进行位置表示，会导致模型泛化性不佳，因为实际应用中会遇到长度大于训练数据集的句子。为解决这个问题，采用正余弦函数来生成位置向量，该方法可对任意长度的句子进行位置编码，具体的嵌入层位置编码向量如式(3)所示：

$$\begin{cases}\boldsymbol{\rho}(pos,2i)=\sin\left(\dfrac{pos}{\sqrt[d_{\text{model}}]{1\,000^{2i}}}\right)\\ \boldsymbol{\rho}(pos,2i+1)=\cos\left(\dfrac{pos}{\sqrt[d_{\text{model}}]{1\,000^{2i}}}\right)\end{cases} \tag{3}$$

3.2.4 主题嵌入

通过吉布斯采样算法对模型参数进行迭代更新，当模型收敛时，我们可以得到每个文档的主题分布，并对其进行归一化，使其和为 1。这样，每个文档都可以用一个 K 维的概率向量表示，其中 K 是主题的个数，向量的每个元素对应一个主题的概率。这个概率数字是一个多维向量，其中每个维度对应一个主题。使用 PCA 算法对高维数据进行降维，将每个文档的 K 维主题向量投影到一个低维空间中，得到一个二维的矩阵 $\boldsymbol{\mu}$。$\boldsymbol{\mu}$ 为 $n\times768$ 的主题嵌入（topic embeddings），表示该段文字的主题信息，用来辅助编码器学习文本的主题信息。主题特征向量的提取方法如算法 1 所示。

算法 1 提取主题向量

输入：文档 D

参数："文档-主题"分布 $\boldsymbol{\theta}$，主题数 K

输出：主题向量 $\boldsymbol{\mu}$

```
M ← LDA(D,K)
centroids,labels ← K-means(M,K)
topic_vectors ← []
for i in range(K):
    indices ← [j for j in range(len(labels)) if labels[j] ==i] and
    rows ← [θ[j] for j in indices] and mean ← average(rows) and
    topic_vectors.append(mean)
μ ← matrix(topic_vectors)
return μ
```

3.3 编码层

编码层由 12 个 Transformer 层叠加构成，其目的是学习文本的深层语义特征，每个 Transformer 层包含一个多头自注意力机制和一个全连接前馈神经网络。多头自注意力子层可以计算任意两个字/词之间的相关性，并将其距离缩减为 1。为解决信息传递过程中记忆偏差问题，将序列依次输入到残差连接和层归一化中。残差连接可以使模型更容易学习到恒等映射，从而降低训练难度。然后将模型的输入与上一层的输出进行加和。层归一化处理可以使模型更加关注差异性信息，提高模型的表达能力，并解决信息传递过程中的梯度爆炸问题。

文本数据首先经过嵌入层转化为序列向量，然后与位置向量相加得到 Transformer 编码器的输入。将输入序列分别通过 3 个线性变换映射为键矩阵（K，keys）、值矩阵（V，value）和查询矩阵（Q，query）输入到多头注意力子层中，计算每个字/词与其他字/词之间的相关权重。多头注意力机制的头数为 h，本模型中每个 Transformer 编码器都含有 12 个多头注意力子层。权重矩阵为 $\boldsymbol{W}^M$。多头注意力层的输出如式(4)所示：

$$MultiHead(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=Concat(head_1,head_2,\cdots,head_h)\boldsymbol{W}^M \tag{4}$$

模型使用 Softmax 函数计算出一个字/词对其他字/词的权重系数，利用编码层学习到的语法结构和语义信息对文本情感进行分类。使用 $\sqrt{d_k}$ 作为稳定模块训练梯度的调和因子，目的是防止内积过大，辅助模型捕捉数据的相关性，解决了传统神经网络的长距离依赖问题。在模型中，自注意力子层的输出如式(5)所示：

$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{5}$$

第一个 Add Norm 层的输出作为前馈神经网络(FFN)的输入，FFN 的输出再经过下一个 Add Norm 层后输出。N_x 是叠加的层数，$\mathrm{ALBERT_{base}}$ 总共使用了 12 层。Transformer 编码器网络可并行抽取文本特征，然后将文本特征通过全连接层映射到样本标记空间，得到的相应的表征向量 $\boldsymbol{T}=\{T_i \mid i\in\{1,2,\cdots,n\}\}$。

编码层的 Transformer 编码器具体结构如图 3 所示。

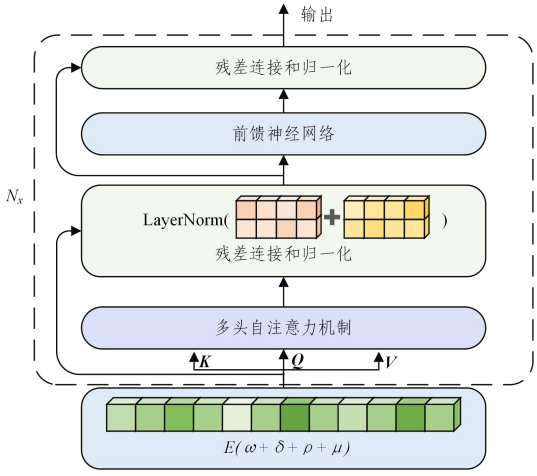


图 3　编码层模块结构

Fig. 3　Coding layer module structure

为了适应细分领域的文本分类任务，本文提出了一种基于 ALBERT 模型的微调方法，即在其顶部添加了一个线性分类器，并用目标领域的数据集进行训练和评估。设 H 为隐藏层的维度，E 为词向量维度，V 为词表大小。ALBERT 模型采用了分解嵌入参数化机制，即通过引入一个中间矩阵来将词汇嵌入层和隐藏层分离，并将原来大小为 $V\times H$ 的词汇嵌入矩阵分解为两个较小的矩阵：一个大小为 $V\times E$ 的词汇嵌入矩阵和一个大小为 $E\times H$ 的中间矩阵。该机制避免了 BERT 中隐藏层和词汇嵌入层之间大小相同但含义不同的问题，从而降低了词汇嵌入层的维度。ALBERT 模型通过分解嵌入参数化机制，将词嵌入维度 E 设为 128，远小于隐藏层维度($H=768$)，并通过一个线性变换映射到隐藏层。这样做可以极大地降低词汇嵌入矩阵的参数量。跨层参数共享是指 ALBERT 模型在所有隐藏层中使用相同的参数，从而显著减少了模型参数量和内存消耗。通过共享机制，中间层的参数量可以忽略，这样既保证了可以得到与 BERT 模型相当甚至更高的准确率，又降低了计算复杂度。

BERT 模型理论总参数量 P_{bert} 的计算过程如式(6)所示：

$$P_{\mathrm{bert}}=V\times H \tag{6}$$

ALBERT 模型理论总参数量 P_{albert} 的计算过程如式(7)所示：

$$P_{\mathrm{albert}}=V\times E+H\times E \tag{7}$$

通过对嵌入层输出的向量表示进行元素求和，得到一个形状为 $n\times 768$ 的二维矩阵 $\boldsymbol{E}=\{E_{\mathrm{Toki}} \mid i\in\{1,2,\cdots,N\}\}$ 作为编码层的输入表示。E_{Toki} 是第 i 个 token 的词嵌入、位置嵌入、段嵌入和主题嵌入的加权和向量；N 为 token 序列的长度，最大不超过 512。这是传递给基于 Transformer 的编码器层的输入表示。向量 $\boldsymbol{E}$ 经过分解嵌入参数化机制映射到隐藏层维度后，作为第一个编码器的输入；每个编码器的输出作为下一个编码器的输入；最后一个编码器的输出是文本中各个字/词融合了全文语义信息后的表征向量 $\boldsymbol{T}$。

3.4　分类层

在编码层后添加平均池化层，对编码层输出的所有 token 嵌入 $\boldsymbol{T}$ 经过平均池化操作，输出固定大小的文档表征嵌入 $\boldsymbol{p}$。该层可以将汉语中的重要词语或短语视为固定的语义组合，提高模型对相邻汉字语义信息的理解能力。具体地，该方法是利用池化操作将过滤器最后一层隐含状态在每个矩阵上的多个特征值中的平均值保留下来，获取到局部区域内的平均特征，从而得到一个句子级别的向量表示 $\boldsymbol{p}$。池化操作的详细方法如算法 2 所示。

算法 2　token 嵌入平均池化算法

输入：n×d 的矩阵 **T**；池化内核大小 m

参数：零填充 g

输出：句子表征 **p**

```
1. p ← zeros(n,768)
2. T_pad ← g(T) and i,j ← 0
3. while i < n
4.    while j < 768 do sum ← 0
5.       for k in range(m)
6.          sum ← sum+T_pad[i+k][j]
7.       p[i][j] ← sum/m and j ← j+1
8.    i ← i+1 and j=0
9. end while
10. return p
```

对池化层的输出进行线性变换，作为全连接层的输入。使用 softmax 函数将逻辑向量转换为概率。将句子表示 $\boldsymbol{p}$ 映射到一个 2 维向量，表示情感极性类别。使用 softmax 函数将输入向量转换为一个[0,1]并且和为 1 的概率分布，以计算情感极性概率。

使用 P 表示输出向量的类别概率。$\boldsymbol{x}$ 表示输入向量，$\boldsymbol{W}_i$ 表示权重矩阵的第 i 行，b_i 表示偏置向量的第 i 个元素，n 表示输出向量的长度。式(8)将输入向量 $\boldsymbol{x}$ 映射到一个长度为 n 的向量上，其中每个元素对应于一个类别，而每个元素的值表示该类别的概率。情感极性概率计算过程如式(8)所示：

$$P=\frac{e^{\boldsymbol{W}_i\boldsymbol{x}+b_i}}{\sum_{j=1}^{n}e^{w_j x+b_j}} \tag{8}$$

4　实验

4.1　数据集的获取与预处理

为了验证本文提出的融合文本主题的特征提升句子的情感分类的有效性，通过 Python 的请求模块向评论内容的 API 发送请求，并将从多个平台获得的 10 万条不重复的酒店、外卖、购物等多领域评论文本组成 gather_senti_100k 数据存储在字典列表结构中。其中，正负向评论各 5 万条，平均长度为 66 个汉字，最大长度 260 个汉字。此外，为充分验证模型的

稳定性，本文在公开的新浪微博评论数据集 weibo_senti_100k 和 online_shopping_10_cats 数据集中重复训练和评估工作。

数据集中标签“0”表示消极情感，“1”表示积极情感。两个数据集字符数量的分布情况如图 4 所示。

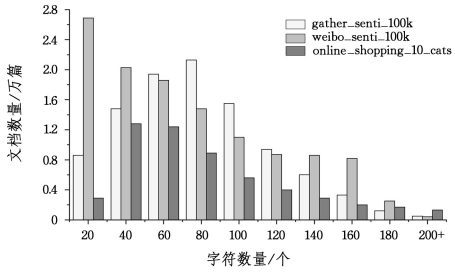


图 4 数据集分布图

Fig. 4 Dataset distribution

训练前，需要对数据集进行分词和去停用词等预处理工作。然后，按照 6∶2∶2 的比例随机抽样各领域数据集，将两个数据集分别划分为训练集、验证集和测试集。这样做的目的是为了保证数据集的代表性和独立性，从而更客观地评估模型的分类效果。

4.2 评测标准

为了评估模型性能，本文引入混淆矩阵的概念，使用准确率(Acc)、召回率(R)和 F1 值(F1)作为模型分类效果的评价指标。混淆矩阵如表 1 所列。

表 1 混淆矩阵

Table 1 Confusion matrix

预测结果	实际结果	
	正向情感	负向情感
正向情感	TP(true positive)	FP(false positive)
负向情感	FN(false negative)	TN(true negative)

1)准确率：不论正负标签，预测正确的样本数($TP+TN$)占样本总数($TP+TN+FP+FN$)的百分比，计算过程如式(9)所示：

$$Acc=\frac{TP+TN}{TP+TN+FP+FN}\times 100\% \tag{9}$$

2)召回率：预测正确并且标签也为正的样本数量(TP)占正样本总数($TP+FN$)的百分比，计算式为：

$$R=\frac{TP}{TP+FN}\times 100\% \tag{10}$$

3)F1 值：因为精确率和召回率经常矛盾，所以使用 F1 值作为调和均值，计算式为：

$$F1=\frac{2PR}{P+R}\times 100\% \tag{11}$$

4.3 环境与参数

本文实验在 NVIDIA 3060 GPU 上进行训练与测试工作，实验平台的具体信息如 2 所列。

表 2 实验环境

Table 2 Lab environment

环境	配置
开发语言	Python 3.7
算法框架	TensorFlow 1.15.0
开发工具	Anaconda3, PyCharm
操作系统	Ubuntu 20.04

模型共有 12 个 transformer 模块、12 个自注意力头、768 个隐藏单元。Adam 是一种基于随机梯度下降(SGD)的一阶式自适应学习率优化算法，通过计算梯度的指数加权移动平均值和平方的指数加权移动平均值来动态地改变学习率。因此，本模型使用 Adam 优化算法来训练模型。参数 Hidden 表示隐藏层的神经元个数。参数 batch_size 是每次训练抓取样本的数量，该参数可以影响模型对文本数据的整体特征的提取和梯度下降的方向。参数 dropout 是一种用于防止神经网络过拟合的正则化技术，是一个介于 0 和 1 之间的浮点数。参数 epochs 是模型在训练过程中遍历整个训练数据集的次数，若准确率在 3 次迭代中都没有提升，则停止训练任务。模型参数设置如表 3 所列。

表 3 模型参数说明

Table 3 Model parameter description

参数	数值
隐藏层激活	Relu
优化器	Adam
最小学习率	1×10^{-5}
Hidden	768
batch_size	32
dropout	0.5
epochs	10
Embedding	128

4.4 实验结果与分析

为验证 LK-ALBERT 模型的效果，本文分别在 3 个数据集上对 4 个经典模型迭代训练 10 次，记录每次迭代的测试集准确率、召回率以及 F1 值，直观反映模型的优劣。选取的对比模型分别为：

1)CNN[12]：使用基于文字的 word2vec 训练词向量，用两个串行卷积层捕捉局部语义特征。在最后一个卷积层的顶部放置一个随时间变化的最大池化层，以选择全局语义特征。使用带脱落的完全连接层来总结特征。

2)LSTM：使用 Word2vec 训练词向量，利用长短期记忆网络提取文本语义信息，随后采用 Softmax 函数判断情感极性。

3)$BERT_{base}$[8]：基于 Google 开源 bert-base 微调，用编码层输出的[CLS]文本聚合序列做分类。

4)$ALBERT_{base}$[9]：基于开源模型微调，将嵌入层提取的文本特征信息输入全连接层后，使用[CLS]进行情感分类。

为了验证本文提出的基于 LDA 结合 K-means 的主题特征提取与池化方法对文本情感分析任务的有效性，本文在多个细分领域的数据集上进行了实验，并与多种方法进行了对比。详细情况如下：

1)L-BERT：将 LDA 输出的主题向量与 BERT 嵌入层的输出向量在输入编码层之前拼接，经过迭代训练后使用[CLS]进行情感分类。

2)LK-ALBERT-cls：基于 albert-base 模型，融合 LDA 和 K-means 算法提取的主题向量，使用[CLS]进行情感分类。

3)LK-ALBERT-max：基于 LK-ALBERT-cls 模型，该模型在最后一层加入最大池化操作，使用过滤器将抽取到的矩阵中多个特征值的最大值保留下来，获取到局部区域内的最突出特征。将池化后的文本特征信息输入全连接层，经过 Softmax 函数实现分类。

4)LK-ALBERT：基于 LK-ALBERT-max 模型，将最大池化替换为平均池化。

5）ERNIE[29]：在每个输入序列的开头添加[CLS]符号，然后将其输入到 ERNIE 3.0 Base 预训练模型中。该模型会对每个符号进行编码，得到其隐藏状态向量。本文取[CLS]符号对应的隐藏状态向量作为整个文本的语义表示，再通过一个全连接层和一个 Softmax 层，得到情绪分类的概率分布。

实验结果表明使用 Adam 的模型性能更出色。在使用 Adam 的算法后，训练集和测试集上的损失（loss）都得到了一定的降低。此外，模型精确度曲线显示测试集的 loss 在前 5 个 epoch 中下降，然后保持稳定。训练集的 loss 在前 5 个 epoch 中下降得更快，然后开始缓慢下降。这表明模型在前 5 个 epoch 中学习了训练数据，并且能够很好地泛化到新数据。但是，在后面的 epoch 中，模型开始过度拟合训练数据，导致测试集的 loss 上升。因此，在算法中加入了早停技术（early stopping），以防止在测试集损失开始上升之后继续训练导致的过度拟合发生。具体情况如图 5 所示。

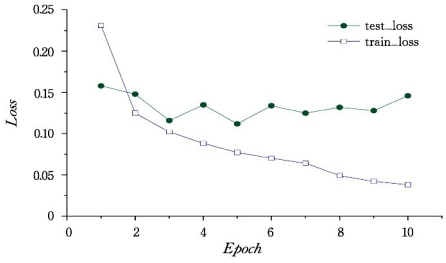


图 5　训练集和测试集的损失变化曲线

Fig. 5　Loss curves of training and test sets

本文通过实验分析了该模型和 5 组对比模型在不同迭代次数下的准确率变化趋势。从趋势图可以看出，该方法相较于其他模型而言，损失函数收敛速度更快，在训练集和测试集上均表现出色。值得注意的是，即使在迭代次数较少时，该模型也能获得较高的分类准确率，这一点是传统深度学习所不具备的优势。与其他模型相比，本模型随着迭代次数的增加，准确率高，收敛速度快且更稳定。在第 8 次迭代后，本模型已经达到了最优参数。值得注意的是，加入平均池化层的模型可以在不增加额外计算成本的前提下提高其有效性。在文本分类任务中，部署平均池化层时，不会增加模型参数，因此不会影响模型的训练。同样，模型每个 eopch 的训练时长也未改变。这一点进一步说明本文提出的方法不仅有效，而且对训练更有益。实验迭代次数与准确率的关系如图 6 所示。

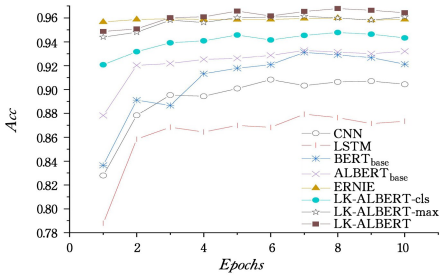


图 6　精确率变化曲线

Fig. 6　Accuracy change curve

由于 BERT 采用创新的双向 Transformer 编码器结构，并且利用大规模中英文数据集进行预训练，有效解决了传统深度学习算法中参数遗忘的问题，同时学习到了丰富的文本语法结构信息，因此基于 BERT 结构的模型分类性能优于基于词嵌入的 LSTM 模型。

本文提出的模型在 ALBERT 的基础上，结合了 LDA 和 K-means 提取的文本主题向量，提高了模型获取情感信息的能力。将本文提出的模型与经典模型 CNN，LSTM，BERT$_{base}$，ALBERT$_{base}$进行对比，结果显示本文提出的模型在 3 个指标均上优于经典模型。通过与 L-BERT，LK-ALBERT-cls，LK-ALBERT-max 模型的对比，证明了本文提出的融合主题特征以及增加池化层的方法有助于提升模型情感分析的效果，同时证明了本文所提模型在文本情感分析领域优于 ERNIE。

所有对比实验都相同实验环境下进行，分类任务训练每间隔 10 steps 评估验证集效果，取验证集最优效果作为汇报指标，具体结果如表 4 所列。

表 4　实验结果

Table 4　Experimental results

模型	gather_senti_100k			weibo_senti_100k			online_shopping_10_cats		
	准确率	召回率	F1 值	准确率	召回率	F1 值	准确率	召回率	F1 值
CNN	0.9083	0.8981	0.9032	0.8839	0.9012	0.8925	0.8982	0.8975	0.8979
LSTM	0.8564	0.8721	0.8643	0.8264	0.8356	0.8310	0.8441	0.8545	0.8493
BERT$_{base}$	0.9312	0.9283	0.9296	0.9286	0.9338	0.9312	0.9292	0.9283	0.9288
ALBERT$_{base}$	0.9328	0.9421	0.9374	0.9262	0.9212	0.9237	0.9306	0.9421	0.9363
ERNIE 3.0base	0.9599	0.9675	0.9637	0.9510	0.9573	0.9541	0.9556	0.9694	0.9625
LK-ALBERT-cls	0.9478	0.9778	0.9626	0.9587	0.9721	0.9654	0.9524	0.9753	0.9637
LK-ALBERT-max	0.9614	0.9645	0.9629	0.9516	0.9680	0.9597	0.9664	0.9663	0.9663
LK-ALBERT	0.9678	0.9875	0.9776	0.9597	0.9698	0.9647	0.9689	0.9809	0.9749

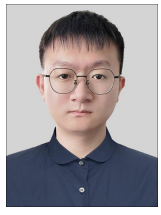
综上所述，本文提出的 LK-ALBERT 模型在情感分析领域相比其他模型在精确率和召回率等评价指标上均有较大的提升。与 BERT 相比，融合了主题特征向量的新算法使准确率提高了 3.67%。

结束语　本文针对文本情感分类问题提出了一种新的方法，在 3 个数据集多个细分领域上对不同模型的性能进行了比较。实验结果表明，融合 LDA 和 K-means 提取的主题特征，能够更有效地捕捉文本语义信息，从而显著提高情感分类的准确性。与以往的 SOTA 模型相比，本文所提模型实现了更稳定、更好的情感分析性能。这些发现为情感分析领域的研究提供了新的思路和方法，有助于相关部门更加准确地掌握公共事件的舆论倾向。未来，将进一步探索如何利用主题信息来增强模型在情感分析任务上的性能和稳定性。

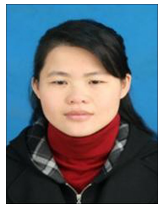
参 考 文 献

[1] WANKHADE M，RAO A C S，KULKARNI C. A survey on sentiment analysis methods，applications，and challenges[J]. Artificial Intelligence Review，2022，55(7)：5731-5780.

[2] TAHERDOOST H, MADANCHIAN M. Artificial Intelligence and Sentiment Analysis: A Review in Competitive Research[J]. Computers, 2023, 12(2): 37.
[3] ZHOU J, YE J M. Sentiment analysis in education research: a review of journal publications[J]. Interactive learning environments, 2023, 31(3): 1252-1264.
[4] LAN Y X, ZHANG L W, WANG H W, et al. Risk-oriented online public opinion abnormal perception and empirical research [J]. Modern intelligence, 2022, 42(3): 102-108.
[5] OSMANI A, MOHASEFI J B, SHI Y. Opinion Mining Using Enriched Joint Sentiment-Topic Model[J]. International Journal of Information Technology & Decision Making, 2023, 22(1): 313-375.
[6] CHATURVEDI I, CAMBRIA E, WELSCH R E, et al. Distinguishing between facts and opinions for sentiment analysis: Survey and challenges[J]. Information Fusion, 2018, 44: 65-77.
[7] ZHOU C, LI Q, LI C, et al. A comprehensive survey on pretrained foundation models: A history from bert to chatgpt[J/OL]. https://arxiv.org/abs/2302.09419.
[8] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [J]. arXiv: 1810.04805, 2018.
[9] LAN Z, CHENM, GOODMAN S, et al. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations[J]. arXiv: 1909.11942, 2019.
[10] TURNEY P D, LITTMAN M L. Measuring praise and criticism: Inference of semantic orientation from association[J]. ACM Transactions on Information Systems (TOIS), 2003, 21(4): 315-346.
[11] DEY S, WASIF S, TONMOY D S, et al. A comparative study of support vector machine and Naive Bayes classifier for sentiment analysis on Amazon product reviews[C]// 2020 International Conference on Contemporary Computing and Applications (IC3A), 2020: 217-220.
[12] CHEN Y. Convolutional neural network for sentence classification[D]. University of Waterloo, 2015.
[13] ALROOBAEAl R. Sentiment Analysis on Amazon Product Reviews using the Recurrent Neural Network(RNN)[J]. International Journal of Advanced Computer Science and Applications, 2022, 13(4): 314-318.
[14] LIN X, CHEN Z Z, WANG Z Q. Attribute-level emotional classification based on unbalanced data and integrated learning[J]. Computer Science, 2022, 49(S1): 144-149.
[15] HU Y L, TONG T Q, ZHANG X Y, et al. In-depth learning emotional analysis method of integrating self-attention mechanism[J]. Computer Science, 2022, 49(1): 252-258.
[16] YAHAV A, VISHWAKARMAI D K. Sentiment analysis using deep learning architectures: a review[J]. Artificial Intelligence Review, 2020, 53(6): 4335-4385.
[17] MIKOLOV T, CHEN K, COEEADO G, et al. Efficient estimation of word representations in vector space[J]. arXiv: 1301.3781, 2013.
[18] VASWANI A, SHAZEER N, PARMAR N, et al. Attention Is All You Need[J]. arXiv: 1706.03762, 2017.
[19] HUANG S C, HAN D H, QIAO B Y, et al. Insumer emotional analysis method based on ERNIE2.0-BILSTM-ATTENTION [J]. Journal of Chinese Computer Systems, 2021, 42(12): 2485-2489.
[20] LIU P, YUAN W, FU J, et al. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing[J]. ACM Computing Surveys, 2023, 55(9): 1-35.
[21] LI M, LI W, WANG F, et al. Applying BERT to analyze investor sentiment in stock market[J]. Neural Computing and Applications, 2020(3): 1-14.
[22] SONG M, LIU Y L. Bert in the application and optimization of the emotional classification of Weibo short text[J]. Journal of Chinese Computer Systems, 2021, 42(4): 714-718.
[23] WANG H, HU X, ZHANG H. Sentiment analysis of com modity reviews based on ALBERT-LSTM[C]// Journal of Physics: Conference Series. Bristol, UK, 2020: 012022.
[24] GAO X, DING G, LIU C, et al. Research on high precision Chinese text sentiment Classification based on ALBERT Optimization[C] // 2023 15th International Conference on Advanced Computational Intelligence(ICACI). Nanjing, China, 2023: 1-6.
[25] BLEI D M, NG A Y, JORDAN M I. Latent Dirichlet Allocation [J]. The Annals of Applied Statistics, 2003, 3: 993-1022.
[26] YAN F, DU T F, MAO J H, et al. Emotional analysis of the stock market text based on emotional dictionary and LDA model [J]. Electronic Measurement Technology, 2017, 40(12): 82-87.
[27] XUE J, CHEN J, HU R, et al. Twitter discussions and emotions about the COVID-19 pandemic: Machine learning approach[J]. Journal of Medical Internet Research, 2020, 22(11): e20550.
[28] BUI Q V, SAYADI K, BUI M. A multi-criteria document clustering method based on topic modeling and pseudoclosure function [C]// Proceedings of the Sixth International Symposium on Information and Communication Technology. Ho Chi Minh City, Vietnam, 2015: 38-45.
[29] SUN Y, WANG S, FENG S, et al. ERNIE 3.0: Large-scale Knowledge Enhanced Pre-training for Language Understanding and Generation [J]. arXiv: 2107.02137, 2021.



YANG Junzhe, born in 1999, postgraduate, is a member of CCF (No. P2586G). His main research interests include sentiment analysis and topic classification.



SONG Ying, born in 1979, Ph.D, postgraduate supervisor, is a member of CCF(No. P2602M). Her main research interests include computer vision and digital twins.