doi:10.3772/j.issn.2095-915x.2024.05.008

基于 BERTopic 主题模型融合 RoBERTa 算法的短文本分类方法研究

刘桂锋 陈亦侯 包翔 韩牧哲

江苏大学科技信息研究所 镇江 212013

摘要：[目的 / 意义] 针对短文本分类中的稀疏问题，提出一种基于 BERTopic-RoBERTa-PCA-CatBoost 模型进行主题概率特征扩展的短文本分类方法。[方法 / 过程] 使用 RoBERTa 模型获取短文本的词向量表示，使用 BERTopic 主题模型提取主题概率特征向量，二者融合进行特征扩展，最后通过 CatBoost 算法分类。[局限] 在分类层面，未使用深度学习算法进行验证；在特征融合层面，未来可以考虑其他的特征融合方法。[结果 / 结论] 提出的 BERTopic-RoBERTa-PCA-CatBoost 模型与 LDA-CatBoost 模型相比在准确率上提升 10.90%，精确率上提升 10.91%，召回率上提升 10.68%。基于主题概率特征扩展的短文本分类方法能够克服单一模型的不足，提高短文本分类的效果。

关键词：短文本分类；词向量；BERTopic 模型；RoBERTa 模型

中图分类号：TP39；G35

Research on Short Text Classification Method Based on BERTopic Topic Modeling and RoBERTa Algorithm

LIU Guifeng CHEN Yihou BAO Xiang HAN Muzhe

Institute of Scientific and Technical Information, Jiangsu University, Zhenjiang 212013, China

Abstract: [Purpose/Significance] To address the sparsity issue in short text classification, this paper proposes a short text classification method based on topic probabilistic feature expansion with BERTopic-RoBERTa-PCA-CatBoost model. [Methods/Processes] The RoBERTa model is employed to obtain word vector representations of short texts. Topic probabilistic feature vectors are extracted using BERTopic topic model, which is then fused with word vectors for feature expansion. Finally, the

基金项目 2024 年江苏省研究生科研与实践创新计划项目“基于 BERTopic 主题概率特征扩展的新闻短文本分类方法研究”（2385）；国家社会科学基金一般项目“科学数据融合模式设计与体系建构研究”（21BTQ080）。

作者简介 刘桂锋（1980-），博士，教授，主要研究方向为科研数据管理、大数据分析、专利分析，E-mail：liuguifeng29@163.com；陈亦侯（1998-），硕士研究生，主要研究方向为机器学习、文本挖掘；包翔（1991-），博士研究生，馆员，主要研究方向为机器学习、文本挖掘；韩牧哲（1990-），博士，讲师，主要研究方向为数字人文、知识管理。

引用格式 刘桂锋 , 陈亦侯 , 包翔 , 等 . 基于 BERTopic 主题模型融合 RoBERTa 算法的短文本分类方法研究 [J]. 情报工程 , 2024, 10(5): 85-98.

doi:10.3772/j.issn.2095-915x.2024.05.008

CatBoost algorithm is utilized for classification. [Limitations] In terms of classification, deep learning algorithms have not been utilized for verification. Regarding feature fusion, future work may consider alternative feature fusion methods. [Results/Conclusions] The proposed BERTopic-RoBERTa-PCA-CatBoost model demonstrates improvements of 10.90% in accuracy, 10.91% in precision, and 10.68% in recall compared to LDA-CatBoost model. The short text classification method based on topic probabilistic feature expansion can overcome the limitations of individual models and enhance the effectiveness of short text classification.

Keywords: Short Textbook Classification; Word Vector; BERTopic Model; RoBERTa Model

引言

近年来短文本数据呈现出一种爆发式的增长态势，例如新闻媒体平台每天都会产生海量的数据，导致这些平台对于信息分类的需求也日益提高。传统的人工数据标注流程耗时过长、质量低下，且受到标注人员主观意识的影响，使得自动化短文本分类方法逐渐取代人工方法，但在处理短文本数据时存在准确率不高、适应性不强等问题。一方面，短文本包含的词语较少，导致传统的文本特征提取方法难以有效捕捉到短文本的语义信息。另一方面，短文本可能生成高维稀疏的特征向量，导致“维数灾难”，增加分类模型的计算复杂度，降低分类效果。基于主题词集等特征扩展方式的短文本分类方法由此受到广泛关注。尽管已有不少研究取得了进展，但此类方法容易导致对短文本主题词的提取不直接、不明确，容易引入噪声等问题。

因此，本研究提出一种基于 BERTopic 主题模型融合 CatBoost 算法的短文本分类方法，具体而言，该方法利用 RoBERTa 模型的文本表示能力来捕获短文本的深层语义信息，经主成分分析降维后能够有效处理高维稀疏矩阵，再通过 BERTopic 主题模型对短文本进行主题建模以发掘短文本中蕴含的潜在主题信息，最后将二者拼接得到的特征向量输入到 CatBoost 分类器进行分类学习。本研究的主要创新之处在于提出一种新的特征提取和特征融合方法，旨在将短文本的主题信息和语义信息有效融入分类模型中，构建出效果更好的短文本分类模型。利用该方法对短文本进行自动分类，不仅有助于信息的有效组织和检索，还可以为信息推荐、情感分析、舆情监控等任务提供支持。

1 相关研究

1.1 传统的短文本表示与分类算法

短文本分类是一个按照一定的分类体系或标准，对短文本数据进行自动化分类的过程。这个任务涉及短文本表示和分类算法两大方面[1]。短文本表示通过提取短文本数据的特征信息，生成包含短文本内部信息的表示，随后将这些表示输入到分类器中按照某种规则进行短文本分类[2]。常见的短文本表示方法包括词袋法如 TF-IDF[3]，词嵌入法如 Word2Vec[4]、Glove[5] 和 FastText[6] 等，基于 Transformer[7]

的多层双向编码器 BERT[8] 的方法以及改进的 BERT 方法如 RoBERTa[9]。常见无监督分类算法包括 K 均值聚类、密度聚类、层次聚类等 [10]。常见带监督分类算法包括支持向量机（SVM）、逻辑回归等 [10]。此外，集成分类算法通过将上述多个弱分类器集成来提高预测精度，比如随机森林算法 [10]、LightGBM[11]、XGBoost[12] 和 CatBoost[13]。

1.2 基于外部语料库的短文本特征扩展方法

短文本数据往往长度较少，信息量不多，这导致它们的特征较为稀疏，使得传统的短文本表示方法若直接结合分类算法，效果并不理想 [14]。为了提高短文本分类算法的分类效果，越来越多的学者尝试通过短文本特征扩展方式扩充语义特征空间 [15]。基于维基百科等外部语料库和知识库的特征扩展方法在一定程度上扩充了信息量，解决了短文本内容不足的问题。张巍琦 [16] 提出了一种基于知识图谱的短文本特征拓展方法。刘懿霆 [17] 提出了基于维基百科的文本样本扩展方法。Wensen 等 [18] 提出了基于 Wikipedia 和 Word2Vec 的特征扩展方法。但是通过外部语料和知识库扩展短文本特征耗时较长，且受限于数据库的质量和专业性，所以部分学者尝试利用基于文本内容自身的特征扩展方法来解决上述问题。

1.3 基于文本自身内容的短文本特征扩展方法

基于文本内容自身的特征扩展方法指的是对短文本内容的词频、主题、语义相关性等属性进行分析挖掘的过程，以往的研究较多集中于从高频词集、主题词集和词汇关联词集等内容出发构建内部语料 [19]。比如，Gu 等 [20] 提出了一种基于关键词扩展的短文本分类算法。李心蕾等 [21] 利用 Word2Vec 和 Sent2Vec 算法生成新浪微博的文本向量化表示形式，来提升分类效果并降低计算成本。曾子明等 [22] 利用 LDA 主题模型提取微博文本主题作为谣言识别模型训练的文档 - 主题特征，并结合用户可信度和微博影响力特征，使用随机森林方法进行谣言识别。唐晓波等 [23] 分别使用 TF-IDF 和 LDA 提取关键词，并利用 Word2Vec 对关键词进行特征扩展，应用于医疗问答社区中的健康问题分类。然而，以往基于短文本主题等内容的特征扩展方法更依赖于人工构建的主题词集质量。主题词集通常针对特定任务进行构建，导致模型在泛化能力方面有所欠缺，也就是说当面对新数据或新任务时，需要重新构建主题词集。

针对上述问题，本文主要采用一种基于主题概率特征扩展的短文本分类方法，设计短文本主题信息和语义信息的特征提取方法，以及将主题特征和语义特征进行拼接的特征融合方法，完成对短文本主题和语义层面特征的扩充。

2 基于主题概率特征扩展的短文本分类的研究思路设计

本研究提出的基于主题概率特征扩展的短文本分类的研究框架如图 1 所示，主要工作分为 3 个部分：（1）设计针对短文本语义信息的特征提取方法；（2）设计针对短文本主题信息的特征提取方法；（3）设计融合主题特征和语义特征的特征向量拼接方法。

doi:10.3772/j.issn.2095-915x.2024.05.008

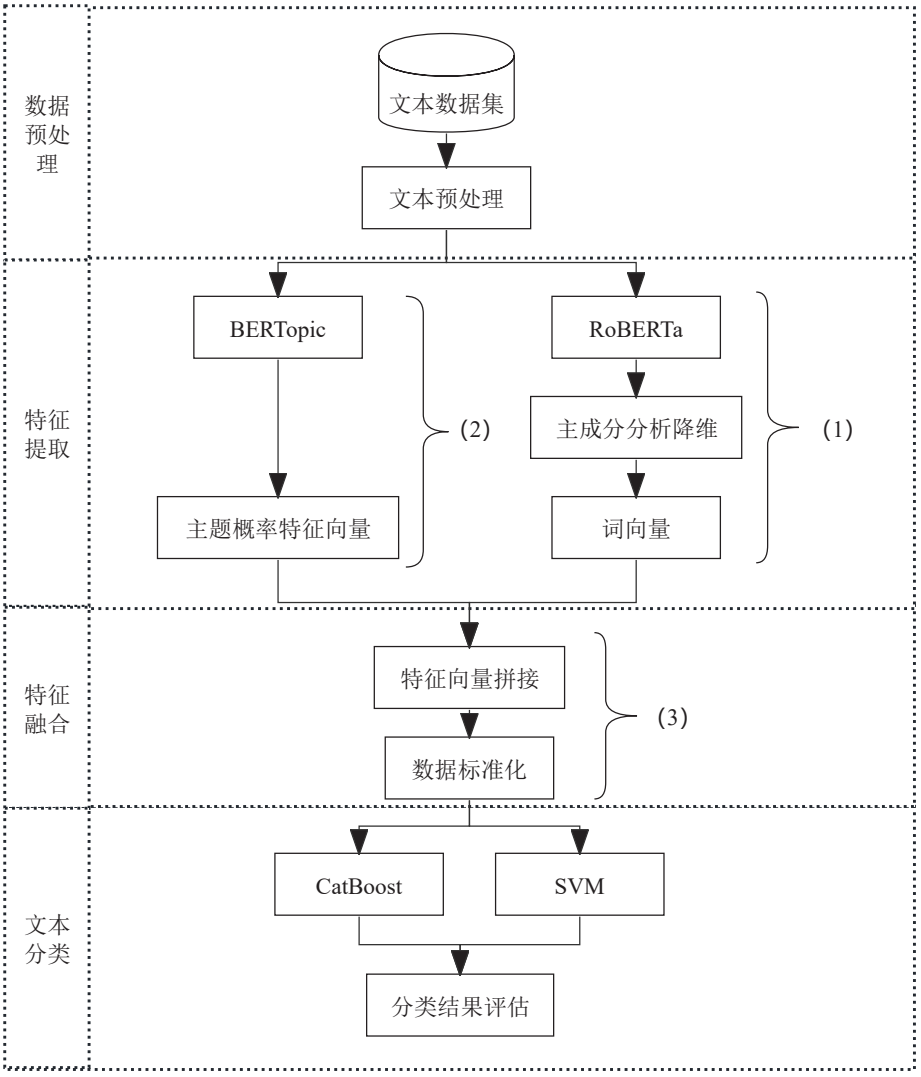

图 1 基于主题概率特征扩展的短文本分类的研究框架

2.1 短文本主题信息和语义信息的特征提取方法设计

本文尝试通过 RoBERTa 预训练模型获取短文本的动态词向量。RoBERTa 预训练模型是一种改进的训练 BERT 模型的方法，在使用 Transformer 进行特征抽取的基础上，结合自注意力机制考虑上下文语义关联，获取动态词向量，从而解决静态词向量无法表示多义词的问题。在此基础上，利用主成分分析将 RoBERTa 模型的输出降维，通过主成分分析法的曲线图确定阈值，确保大部分特征信息得到保留。主成分分析法降维对于下个阶段的特征融合来说确保了主题特征和语义特征的相对平衡，对于分类模型来说则去掉了噪点信息，能够大幅度减少模型的训练时间并且提高模型的效果。针对短文本语义信息的特征提取方法如图 2 所示。

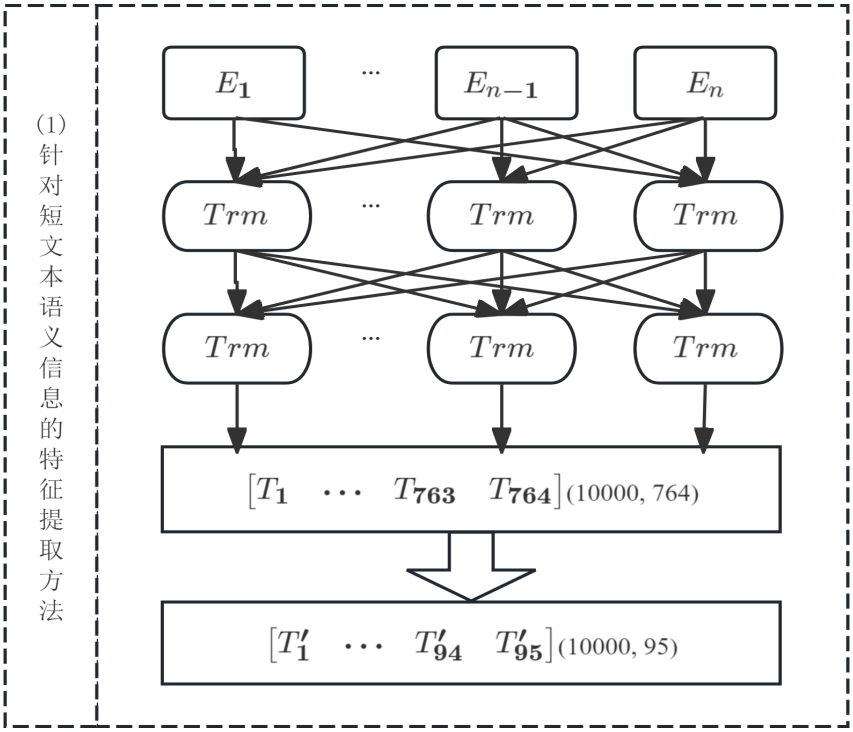

图 2 短文本语义信息的特征提取方法

其中 E_1, …, E_{n-1}, E_n 为词嵌入层的输入，*Trm* 为 Transformer 编码器，[T_1, …, T_{763}, T_{764}] 为 RoBERTa 模型输出的短文本语义特征向量组成的特征矩阵，（10000, 764）是矩阵维度；[T'_1, …, T'_{94}, T'_{95}] 为经过主成分分析降维后的特征矩阵，（10000, 95）是矩阵维度。

本文尝试使用 BERTopic[24] 模型获取短文本的主题向量。首先，通过 BERT 或其他嵌入技术将文档转换为向量表示；再对 BERT 模型处理好的数据使用 UMAP 算法降维后通过 HDBSCAN 算法进行无监督的聚类，以将文档划分为不同的类别；最后结合 C-TF-IDF 方法提取主题词汇，以生成紧密相关的主题群，最后使用模型输出的文档主题概率特征向量来表征短文本。短文本主题信息的特征提取方法如图 3 所示。

其中，BERT 层嵌入文档，UMAP 和 HDBSCAN 层聚类文档，C-TF-IDF 层生成主题表示；[$topic_1$, …, $topic_5$, $topic_6$] 为 BERTopic 模型输出的短文本主题特征向量组成的特征矩阵，（10000, 6）是矩阵维度。

2.2 融合主题特征和语义特征的特征向量拼接方法设计

本文所提出的特征融合方法是使用 BERTopic 主题模型提取出的主题概率特征集合 [$topic_1$, …, $topic_5$, $topic_6$]，结合 RoBERTa 预训练模型提取的经主成分分析降维后的词向量集合 [T'_1, …, T'_{94}, T'_{95}] 进行特征融合，达到一种信息互补的效果。为了得到一个融合多粒度特征的文本表示，通过特征拼接的方式融合 2 种类型的特征，进而得到文本的多粒度特征 F，如式（1）所示。本文选择的特征拼接方法是一种无参数的方法，它不需要预先设定融合权重或进行复

doi:10.3772/j.issn.2095-915x.2024.05.008

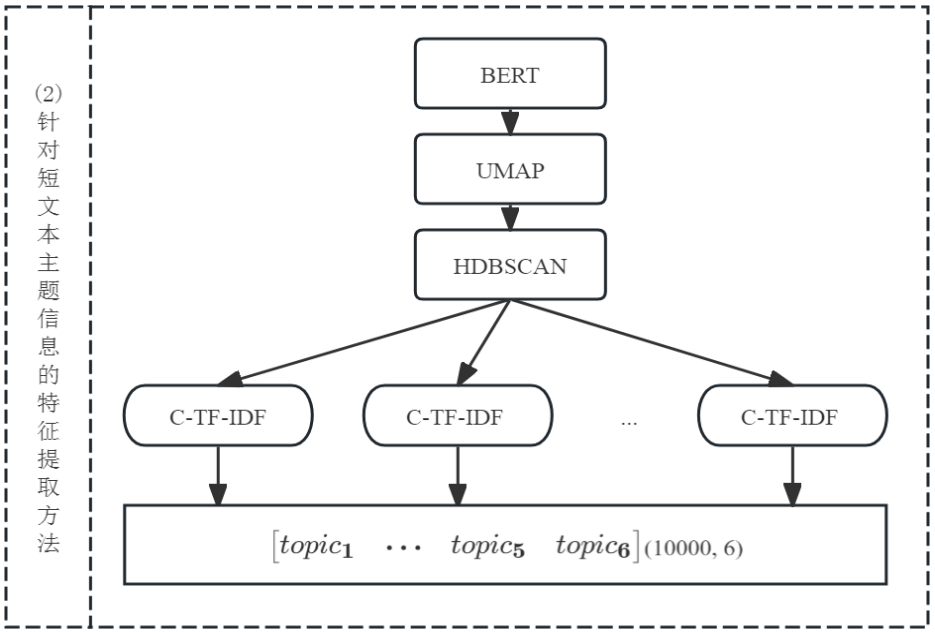

图 3　短文本主题信息的特征提取方法

杂的优化过程，因此更容易实施且能够灵活地适应不同的特征类型和数量。由于这两组特征经过不同的处理，尺度和分布可能会有所不同。为确保模型能够平等地考虑每个特征，对拼接后的特征向量进行归一化处理。

$$F=[topic_1, \ldots, topic_5, topic_6, T'_1, \ldots, T'_{94}, T'_{95}] \quad (1)$$

其中，$[topic_1, \ldots, topic_5, topic_6]$ 为 *BERTopic* 模型输出的短文本主题特征向量组成的主题特征矩阵；$[T'_1, \ldots, T'_{94}, T'_{95}]$ 为经过主成分分析降维后的 *RoBERTa* 模型提取的短文本语义特征向量组成的语义特征矩阵。

3　实验过程

3.1　数据来源以及数据预处理

本文选择清华大学中文文本语料库 THUCNews[25] 数据集作为数据来源，从中抽取财经、家居、教育、科技、社会、时尚、时政、体育、游戏、娱乐共 10 个分类类别，每个类别分别抽取 1000 条样本，共 10000 个样本，按照 8:2 的比例划分训练集和测试集。

新闻短文本数据集的主成分分析 2D 投影图如图 4 所示，图中 x 轴和 y 轴分别代表第一主成分和第二主成分，也就是最大化保留数据特征的两个主成分方向。类别 9（娱乐）在图中呈现出较强的聚集性，对数据点的聚类起到关键作用，而类别 2（教育）等较为分散，对聚类的影响较小。

本实验所用的环境为 Python3.9，数据预处理环节使用 Python 中的 pandas2.2.0 包对数据进行初步的清洗去重后，加载哈尔滨工业大学停用词表、四川大学机器智能实验室停用词库等多个停用词词典，使用 Python 中的 jieba0.42.1 包对摘要文本进行分词，并去除文本中所包含的停用词和其他无意义字符。主题建模环节主

要通过 bertopic0.16.0 和 gensim4.3.2 包来进行模型训练与输出；分类器构建环节选用 scikit-learn1.4.0、catboost2.0 包；结果可视化呈现通过 matplotlib3.7.2 包来实现。

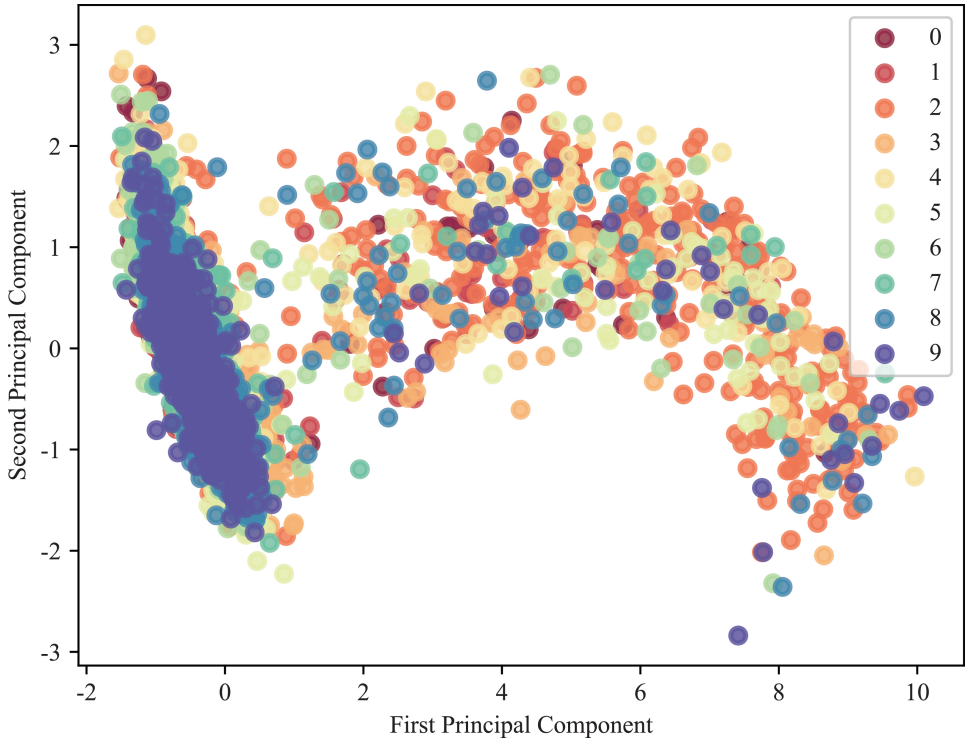

图 4 新闻短文本数据集的主成分分析 2D 投影图

3.2 主题模型参数设置

BERTopic 模型的具体参数如下：嵌入文档，使用中文文本的预训练词向量模型“bert_base_chinese”；初始化 UMAP 和初始化 HDBSCAN 过程使用默认参数；主题数设为 nr_topics =“auto”，在进行主题提取时不限定主题数，这是因为 BERTopic 模型在建模过程中会自动选择最佳参数。

Top2vec 模型的具体参数如下：嵌入文档，使用针对多语言（包括中文）的预训练模型“distiluse-base-multilingual-cased”；设置 min_count=10，表示在构建共现矩阵时，一个词语至少需要出现在 10 个文档中才会被考虑；设置模型训练所需的速度 speed=“learn”，平衡训练速度和训练质量；指定用于并行训练的线程数 workers=8，加快模型训练速度。

LDA 主题模型困惑度的计算通过模型训练时返回的概率分布矩阵得出，LDA 模型在不同主题数下的困惑度指标，如图 5 所示。根据结果选择困惑度最小的主题个数作为模型主题数，LDA 模型最优主题个数 Optimal number of topics = 29；random_state = 42 控制随机数生成的种子，用于确保结果的可重复性。其他参数设置为默认参数。

doi:10.3772/j.issn.2095-915x.2024.05.008

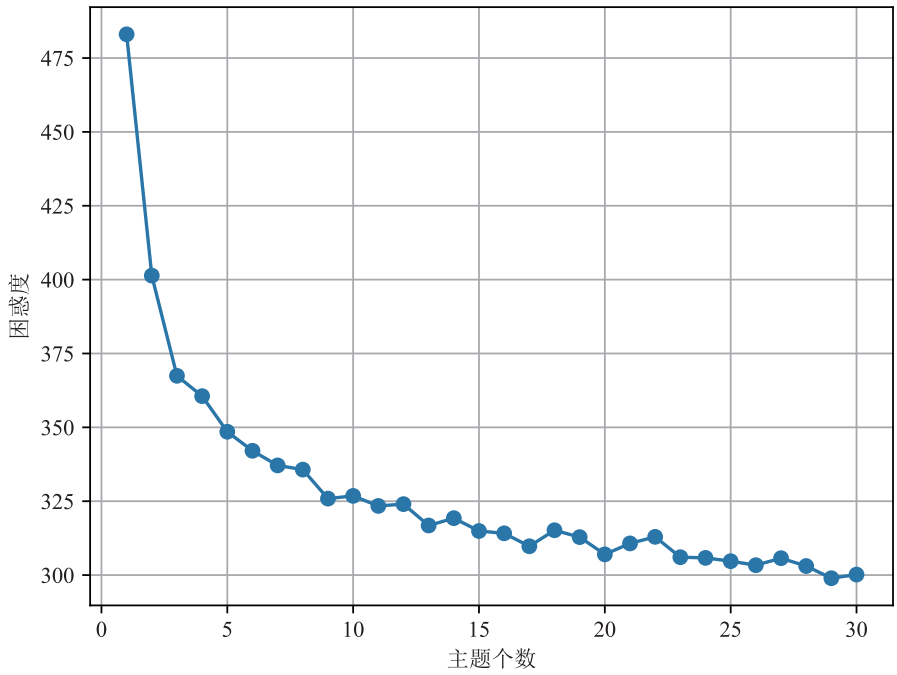

图 5 LDA 模型的困惑度

3.3 其他参数设置

图 6 是主成分分析的方差解释率图。其中每个点表示一个主成分，y 轴值表示该主成分对总方差的解释比例。随着主成分数量的增加，累积方差解释率通常会逐渐增加，但增长速度会逐渐放缓。曲线的拐点通常表示增加更多主成分对总方差的解释增加不再显著。因此，通过选择曲线拐点附近的点，将参数设置为要保留的方差百分比 95% 作为阈值。

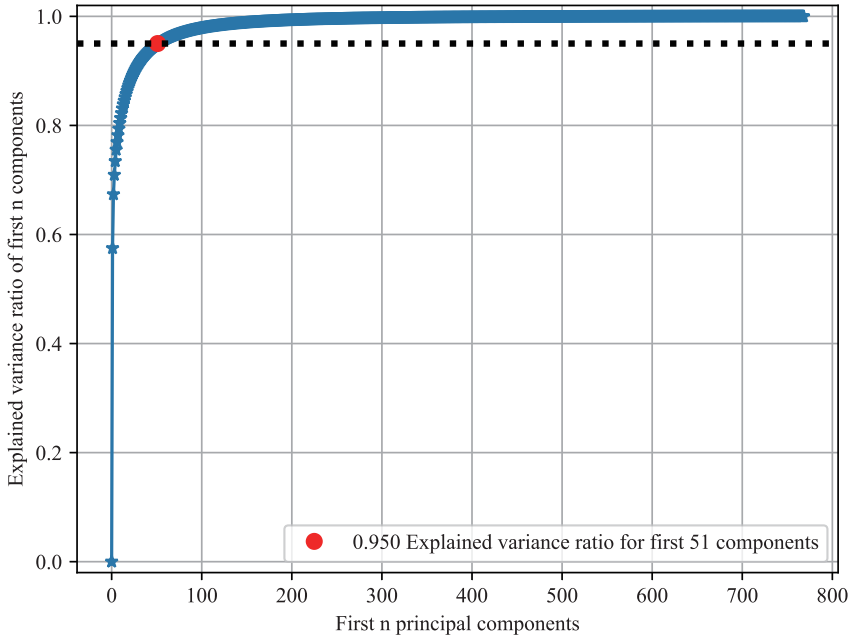

图 6 主成分分析的方差解释率图

CatBoost 分类器参数设置如下：iterations 为迭代次数，默认为 1000；learning_rate 表示模型的学习率，设置 RoBERTa-CatBoost（降维前）为 0.087382，设置其他模型为 0.087979；其余参数均使用默认参数。

4 实验结果分析

4.1 词向量模型与主题模型融合的有效性实验

为验证文本特征融合方法的有效性，将其与前人的分类方法进行对比。由于 LDA[26]、Top2vec[27] 作为经典的主题模型方法被广泛应用，TF-IDF 模型是传统的文本向量表征方式之一，同时支持向量机（SVM）等经典机器学习算法在各种分类任务中展现出优越的性能。因此，分别设置 LDA-CatBoost 和 LDA-SVM 为使用 LDA 模型所表征的文本特征向量，结合 CatBoost 和 SVM 分类器进行分类对比，也就是刘爱琴等 [28] 提出的主题模型融合分类算法的文本自动分类方法。分别设置 Top2vec-CatBoost 和 Top2vec-SVM 为使用 Top2vec 模型所表征的文本特征向量，结合 CatBoost 和 SVM 分类器进行分类对比。分别设置 TF-IDF-CatBoost 和 TF-IDF-SVM 为使用 TF-IDF 模型所表征的文本特征向量结合 CatBoost 和 SVM 分类器进行分类对比。设置 BERTopic-RoBERTa-CatBoost 为文本所提出的方法，即结合经 PCA 降维后的 RoBERTa 模型的词向量表示与 BERTopic 模型的主题向量表示进行特征向量拼接，输入到 CatBoost 分类器进行分类。在 THUCNews 新闻数据集上运用上述方法得到的测试结果的准确率、精确率、召回率、F1 值、AUC 值比较结果如表 1 所示。

表 1 对比实验结果

模型	准确率	精确率	召回率	F1 值	AUC 值
LDA-CatBoost	0.7455	0.7469	0.7480	0.7455	0.9615
LDA-SVM	0.7350	0.7346	0.7375	0.7350	0.9499
Top2vec-CatBoost	0.6885	0.6900	0.6873	0.6866	0.9395
Top2vec-SVM	0.6675	0.6803	0.6660	0.6659	0.9324
TF-IDF-CatBoost	0.6535	0.7384	0.6546	0.6727	0.9325
TF-IDF-SVM	0.6040	0.7685	0.6077	0.6359	0.9406
BERTopic-RoBERTa-PCA-CatBoost	**0.8545**	**0.8560**	**0.8548**	**0.8550**	**0.9772**
BERTopic-RoBERTa-PCA-SVM	0.7840	0.7897	0.7850	0.7865	0.9646

对表 1 中的实验结果进行分析得出如下结论：

本文所提出的 BERTopic-RoBERTa-PCA-CatBoost 模型相对于 LDA-CatBoost 模型在分类的准确率上提升 10.90%，精确率上提升 10.91%，召回率上提升 10.68%，说明本文方法能够解决短文本分类中的稀疏问题，进一步提升分类效果。实验结果表明相较于支持向量机模型，结合集成算法 CatBoost 模型的整体分类效果更高，这说明了

doi:10.3772/j.issn.2095-915x.2024.05.008

使用集成学习算法构建分类器更适用于本文所提出的特征融合场景。根据实验结果，选择效果最优的 BERTopic-RoBERTa-PCA-CatBoost 作为本文的最终分类模型。

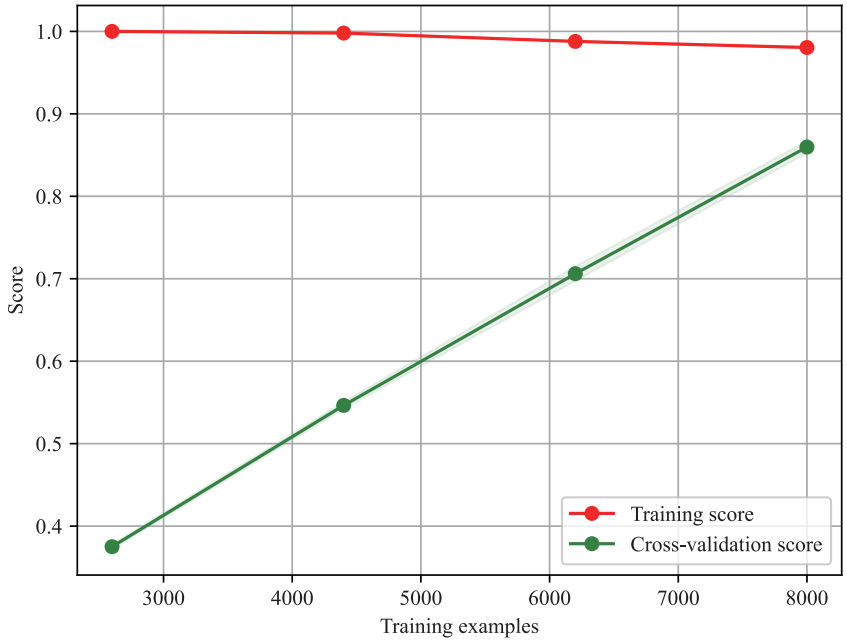

图 7 BERTopic-RoBERTa-PCA-CatBoost 模型的学习曲线图

BERTopic-RoBERTa-PCA-CatBoost 模型的学习曲线图如图 7 所示，随着样本数量的增加，模型在测试集上的交叉验证分数呈明显上升趋势，说明模型的效果在稳步提升，并且在训练完所有样本后达到最高点。另外，随着样本数量的增加，模型在训练集上的分数略下降，说明过拟合的风险降低了。这意味着模型不再过度拟合训练数据中的噪声或细节。从图 7 中可以推断出，若继续扩大样本数量，该模型的效果将进一步提升。

BERTopic-RoBERTa-PCA-CatBoost 模型在短文本分类任务中的优势主要来自三个方面：主题模型 BERTopic 更能有效地捕捉文本中的主题信息；RoBERTa 作为预训练模型，其强大的语义表示能力使得文本的动态上下文语义特征得以充分提取；而 CatBoost 作为分类器，则能基于融合特征进行准确分类。在新闻短文本分类的实际应用中，BERTopic-RoBERTa-PCA-CatBoost 模型提高了分类的准确性。

BERTopic-RoBERTa-CatBoost 模型的 ROC 曲线图以及精确率 — 召回率曲线图分别如图 8、9 所示。精确率 — 召回率曲线展示了不同类别数据在精确率和召回率这两个维度上的综合表现，图 9 中曲线整体下降的趋势明显，因为精准率和召回率是两个相互制约的指标，随着精准率逐渐增大召回率会逐渐地减小，曲线拐点位置是精准率和召回率的平衡位置。从图 8 和图 9 中可以看出，类别 0（财经）、类别 3（科技）和类别 9（娱乐）的曲线较为理想，表明分类器在这些类别上分类效果和性能较好。而类别 8（游戏）的曲线则相对较低，说明分类器在该类别上分类效果和性能均有待提升。

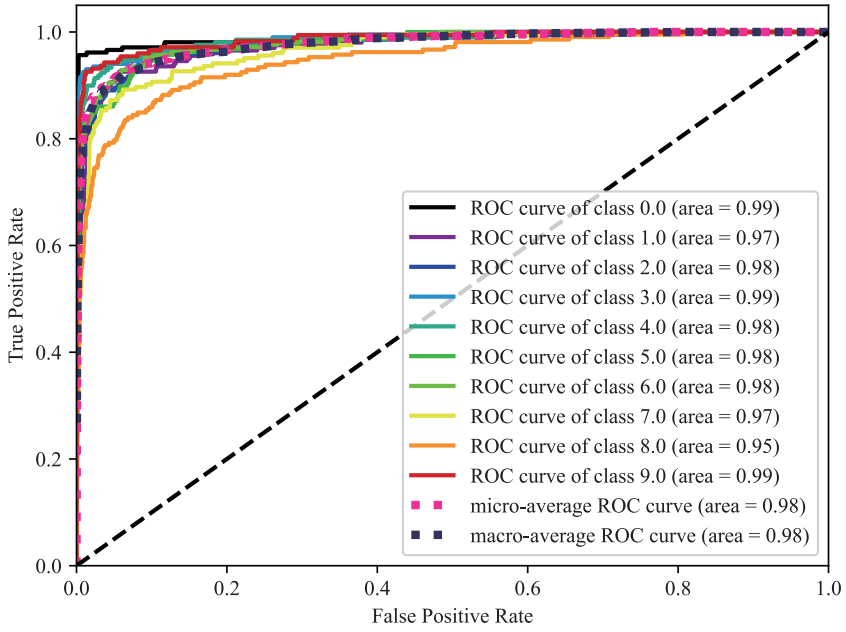

图 8 BERTopic-RoBERTa-CatBoost 模型的 ROC 曲线图

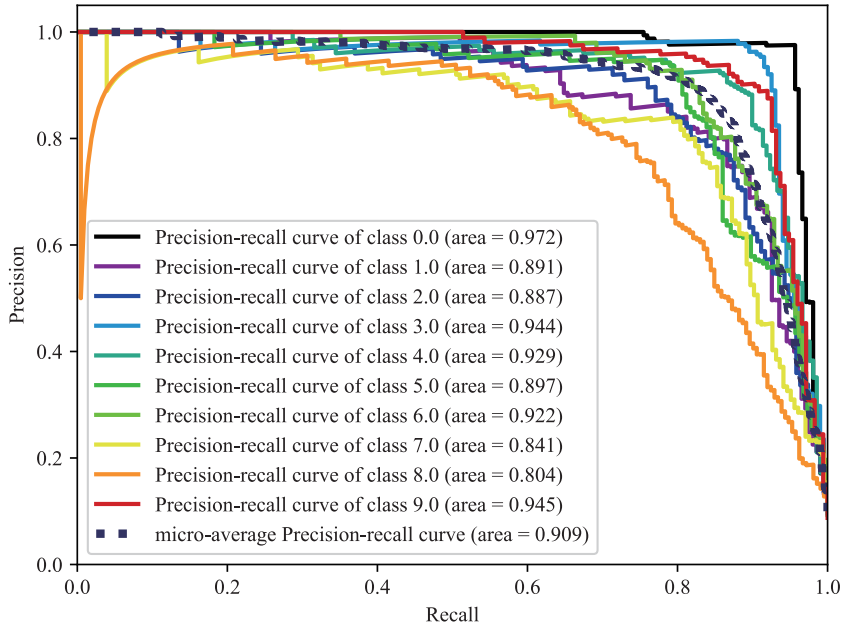

图 9 BERTopic-RoBERTa-CatBoost 模型的精确率 - 召回率曲线图

4.2 基于BERTopic-RoBERTa-PCA-CatBoost模型的消融实验

为了验证 BERTopic-RoBERTa-PCA-CatBoost 模型中各组件的有效性，并评估组件对整体性能的贡献，对模型进行消融实验。设置 RoBERTa 为去除主题模型 BERTopic 和 PCA 模块的模型结构；设置 RoBERTa-PCA-CatBoost 为去除主题模型 BERTopic 的模型结构；设置 BERTopic-CatBoost 为去除预训练模型 RoBERTa 模块的模型结构。基于 BERTopic-RoBERTa-PCA-CatBoost 模型的消融实验比较结果如表 2 所示。

doi:10.3772/j.issn.2095-915x.2024.05.008

表 2　消融实验比较结果

模型	准确率	精确率	召回率	F1 值	AUC 值
RoBERTa-Catboost	0.6340	0.6344	0.6352	0.6328	0.9244
RoBERTa-PCA-CatBoost	0.6235	0.6275	0.6252	0.6247	0.9175
BERTopic-CatBoost	0.8155	0.8398	0.8178	0.8183	0.9687
BERTopic-RoBERTa--PCA-CatBoost	**0.8545**	**0.8560**	**0.8548**	**0.8550**	**0.9772**

从表 2 中可以看出，RoBERTa-PCA-CatBoost 模型比 RoBERTa-Catboost 模型在准确率上降低 1.05%，在 F1 得分方面降低 0.81%，在 AUC 值上降低 0.69%。这说明主成分分析成功保留了大部分的特征，对分类模型效果的影响很小。利用主成分分析法降维的原因在于 RoBERTa 模型输出的特征数量远超主题模型，如果直接进行特征拼接的话，BERTopic 模型析出的主题信息可能被忽略，起不到特征融合的意义。因此，加入主成分分析这一特征提取方法是进行有效特征融合的前提，使分类模型更为均衡地考虑到各个层面的特征，从而提升分类效果。

实验结果表明本文提出的 BERTopic-RoBERTa-PCA-CatBoost 模型相较于 BERTopic-CatBoost 模型在准确率、召回率、F1 得分上分别提升了 3.90%、3.67% 和 0.85%；相较于 RoBERTa-PCA-CatBoost 模型的方法在准确率、召回率、F1 得分上分别提升了 23.10%、23.03% 和 5.97%，验证了本文所提出的词向量模型与主题模型融合方法的有效性。由于新闻短文本篇幅有限，文本中词汇的共现模式较为显著，分类时往往需要对文本进行深层次的语义理解。主题模型如 BERTopic 能够有效地捕捉这些共现模式，所以分类效果相对较优；而依赖文本表示的分类方法在处理具有复杂语义的新闻短文本时，可能对语义理解得不够充分，因此分类效果相对较差。本文使用主题模型 BERTopic 提取无监督聚类结果，结合词向量模型 RoBERTa 经主成分分析降维后做了特征融合实现的监督分类算法，二者融合达到了信息互补的效果，因此效果最优。

5　结论

本文主要提出一种基于主题概率特征扩展的短文本分类方法，通过设计短文本主题信息和语义信息的特征提取方法，以及融合主题特征和语义特征的特征向量拼接方法，完成对短文本主题和语义层面的特征扩展，省去了人工构建主题词集的人力成本。这为情报学提供了一种新的文本分类视角和理论支持。同一文本数据往往蕴含着多重维度的特征分布，包括语义特征、结构特征、情感特征等，特征的多样性和丰富性使得特征融合技术尤为重要。通过本文所提出的特征提取和特征融合方式能够更全面、更准确、更有效地整合多种类型的特征，在情报学信息推荐、信息检索、情感分析、舆情监控等领域具有广泛的应用前景。

实验结果表明，融合主题特征向量和词向

量能够对短文本分类效果产生正向影响，这在一定程度上克服了短文本内容的稀疏问题，提高分类的准确率。此外，引入主成分分析降维不仅有助于特征融合，还能够大幅缩短分类器的迭代时长，提升分类效率。未来可以通过扩大数据集、探索更多的特征提取方法和分类算法，尝试将本文提出的短文本自动分类技术应用于更多的情报分析场景，包括：（1）在分类层面，未来的研究中还可以尝试使用深度学习模型如卷积神经网络 CNN、循环神经网络 RNN、长短期记忆网络 LSTM 等分类器进行验证，比较和评估该方法结合不同的分类算法的分类准确率和性能；（2）在特征融合层面，未来可以考虑其他的特征融合方法，如 TextCNN 的融合门机制等方式。

参 考 文 献

[1] GE J W, LIN S C, FAN Y Q. A Text classification algorithm based on topic model and convolutional neural network[J]. Journal of Physics: Conference Series, 2021, 1748(3): 32-36.

[2] 许和旭 , 王兰成 . 基于语料库文本自动分类算法及应用比较研究 [J]. 图书情报导刊 , 2021, 6(6): 45-53.

[3] JONES K S. Index term weighting[J]. Information storage and retrieval, 1973, 9(11): 619-633.

[4] MIKOLOV T. Efficient estimation of word representations in vector space[J]. arxiv preprint arxiv:1301.3781, 2013.

[5] PENNINGTON J, SOCHER R, MANNING C D. Glove: Global vectors for word representation[C]// Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014: 1532-1543.

[6] JOULIN A, GRAVE E, BOJANOWSKI P, et al. Fasttext. zip: Compressing text classification models[J]. arXiv preprint arXiv:1612.03651, 2016.

[7] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017, 12: 6000-6010.

[8] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[9] ADOMA A F, HENRY N M, CHEN W. Comparative analyses of bert, roberta, distilbert, and xlnet for text-based emotion recognition[C]//2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP). IEEE, 2020: 117-121.

[10] 周志华 . 机器学习 [M]. 北京 : 清华大学出版社 , 2006: 73-224.

[11] KE G, MENG Q, FINLEY T, et al. Lightgbm: A highly efficient gradient boosting decision tree[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017: 3149-3157.

[12] CHEN T, GUESTRIN C. Xgboost: A scalable tree boosting system[C]//Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016: 785-794.

[13] HANCOCK J T, KHOSHGOFTAAR T M. CatBoost for big data: an interdisciplinary review[J]. Journal of big data, 2020, 7(1): 94.

[14] GE J W, WANG H X, FANG Y Q. Short Text Classification Method Combining Word Vector and WTTM[C]// IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference. Chongqing: IEEE, 2020: 1994-1997.

[15] 王连喜 . 微博短文本预处理及学习研究综述 [J]. 图书情报工作 , 2013, 57(11): 125-131.

[16] 张巍琦 . 基于知识图谱的短文本分类研究 [D]. 成都 : 电子科技大学 , 2020.

[17] 刘懿霆 . 基于维基百科的文本样本扩展方法及其应用研究 [D]. 上海 : 上海大学 , 2019.

[18] WENSEN L, ZEWEN C, JUN W, et al. Short text classification based on Wikipedia and Word2vec[C]//2016 2nd IEEE International

doi:10.3772/j.issn.2095-915x.2024.05.008

Conference on Computer and Communications (ICCC). IEEE, 2016: 1195-1200.

[19] 邵云飞 , 刘东苏 . 基于类别特征扩展的短文本分类方法研究 [J]. 数据分析与知识发现 , 2019, 3(9): 60-67.

[20] GU Y, SHEN J. Short text classification based on keywords extension[C]//2019 Chinese Automation Congress (CAC). IEEE, 2019: 2616-2621.

[21] 李心蕾 , 王昊 , 刘小敏 , 等 . 面向微博短文本分类的文本向量化方法比较研究 [J]. 数据分析与知识发现 , 2018, 2(8): 41-50.

[22] 曾子明 , 王婧 . 基于 LDA 和随机森林的微博谣言识别研究——以 2016 年雾霾谣言为例 [J]. 情报学报 , 2019, 38(1): 89-96.

[23] 唐晓波 , 高和璇 . 基于关键词词向量特征扩展的健康问句分类研究 [J]. 数据分析与知识发现 , 2020, 4(7): 66-75.

[24] EGGER R, YU J. A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts[J]. Frontiers in sociology, 2022, 7: 886498.

[25] 孙茂松 , 李景阳 , 郭志芃 , 等 . THUCTC：一个高效的中文文本分类工具包 [EB/OL]. (2016-01-01) [2023-12-31]. http://thuctc.thunlp.org/.

[26] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3: 993-1022.

[27] ANGELOV D. Top2vec: Distributed representations of topics[J]. arXiv preprint arXiv:2008.09470.

[28] 刘爱琴 , 郭少鹏 , 张卓星 . 基于 LDA 模型融合 Catboost 算法的文本自动分类系统设计与实现 [J]. 国家图书馆学刊 , 2023, 32(5): 84-92.