基于SVM-RFE与Transformer-TBAM的高校邮件分析研究

李振[1]，李智超[2]，陈琳[1]
（1.山东大学信息化工作办公室，山东 济南 250100；2.山东大学（威海）信息化工作办公室，山东 威海 265209）

摘　要：通过挖掘高校电子邮件文本数据并进行分析，可以帮助教职工更好地了解学生的意见和建议，提高管理效率。目前，深度学习方法是文本情感分析的主要方法，然而现有的方法没有充分利用中文文本中的特征。为解决此问题，提出基于SVM-RFE与Transformer-TBAM架构模型处理高校邮件，该架构重构了双通道注意力模型及特征筛选机制以深度提取有效特征信息。实验表明，该算法在高校邮件数据集分类效果达到了94.67%的准确率，比传统算法高出1.2%。

关键词：SVM；高校邮件；Transformer；注意力机制
中图分类号：TP391
文献标志码：A
DOI：10.11959/j.issn.1000–436x.2024229

Research on university email analysis based on SVM-RFE and Transformer-TBAM

LI Zhen[1], LI Zhichao[2], CHEN Lin[1]
1. Informatization Office, Shandong University, Jinan 250100, China
2. Informatization Office, Shandong University (Weihai) , Weihai 265209, China

Abstract: By mining and analyzing email text data from universities, it can help faculty members better understand students’opinions and suggestions, and improve management efficiency. At present, deep learning methods are the main approach for text sentiment analysis, but existing methods have not fully utilized the features in Chinese text. To address this issue, a framework based on SVM-RFE and Transformer models was proposed for processing university emails. This architecture reconstructs a dual branch attention model and feature filtering mechanism to deeply extract effective feature information. The experiment shows that the algorithm achieves an accuracy of 94.67% in the classification of university email datasets, which is 1.2% higher than traditional algorithms.

Keywords: SVM, college email, Transformer, attention mechanism

0　引言

随着互联网技术的持续演进，高校与学生的沟通方式发生了显著变革。电子邮件已成为一种核心沟通手段，在学校管理、教学及科研活动中扮演着举足轻重的角色。高校电子邮件内容丰富多样，蕴含从积极鼓励到中性描述乃至消极批评等多种情感信息。这些情感信息深刻影响着学生，进而影响了学生与学校之间的关系。因此，深入分析并识别大学电子邮件的内容十分重要。近年来，神经网络及深度学习技术在文本分析领域取得了显著成就，尤其是Transformer模型[1]，凭借其强大的自注意力机制和位置编码策略，成为自然语言处理研究的热点。然而，传统Transformer模型易陷入过拟合困境，在文本分类任务中往往难以精准捕捉情感信息。为应对这些挑战，本文提出了一种基于改进Transformer模型的大学电子邮件文本情感增强分

收稿日期：2024–10–22
通信作者：陈琳，chenlin@sdu.edu.cn

析方法，对Transformer的自注意力机制进行了优化，旨在提升其在文本情感分类任务中的准确性与泛化能力。

为验证所提方法的有效性，本文在2个公开数据集及高校电子邮件上进行了实验。实验结果显示，相较于传统Transformer模型，本文方法在情感分类任务中展现出更优的性能与泛化能力。通过采用多分支注意力机制深入挖掘文本信息，并改进了从各子空间中提取信息的方式，本文方法显著提升了文本分类的有效性。

本文的主要贡献如下。

1) 分别利用Word2Vec[2]与Glove模型[3]训练词向量，并通过支持向量机（SVM）-递归特征消除（RFE）方法对部分词向量进行特征选择，随后送入Transformer模型进行训练。经SVM-RFE筛选后的词向量特征更具代表性。

2) 提出了Transformer-TBAM（double branch attention model）文本情感分类框架。该方法融合了注意力机制与卷积神经网络的概念，利用多分支通道提取细粒度的文本信息特征，以捕获更有效的信息。在大学电子邮件文本数据集和公共数据集上的对比实验表明本算法准确率上显著优于多种同类算法。

1　研究背景

文本分类作为自然语言处理领域的一个重要研究方向，自1957年起奠定了其研究基础。随后，深度学习技术的广泛普及极大地促进了基于该技术的文本分类方法的研究进展，其中涉及了卷积神经网络（CNN）[4]、循环神经网络（RNN）[5]以及长短期记忆（LSTM）[6]等多种模型的应用。2023年，Umer等[7]提出了一种利用FastText词嵌入生成n-gram向量的方法，该方法能有效应对未登录词问题。与此同时，Polat等[8]则探索了使用Glove模型结合BiLSTM进行假新闻分类的新途径。针对文本特征提取不充分及准确率不高的问题，Li等[9]设计了一种双通道中文文本分类模型，显著增强了传统方法的效果。

2　模型算法及实验步骤

本文模型使用Word2Vec和Glove算法分别获取词向量，之后使用SVM-RFE算法获取后Glove算法的词向量有效特征，通过向量扩展方法，集成最终的词向量。将获取到的融合词向量送入Transformer-TBAM模型训练，最终获取最终分类结果。模型整体架构如图1所示。

2.1　文本预处理

首先，使用Word2Vec和Glove模型训练大学电子邮件文本以获得单词嵌入，这些词嵌入最终合并到一个融合词向量中。预处理的数据集$\boldsymbol{X}=\{x|x_1,x_2,\cdots,x_i,\cdots,x_N\}$包含$N$个数据点，其中$x_i=\{x_{i1},x_{i2},\cdots,x_{iM}\}$表示单个词嵌入，$M$表示词嵌入的特征维度。词向量计算模块是基于对数据集$\boldsymbol{X}$的一系列计算来获得合并特征字典$\boldsymbol{D}$的过程。$\boldsymbol{X}$被送到词向量计算模型中，该模型由Word2Vec与Glove组成。$\boldsymbol{X}$经过模型训练来生成M维特征的权重矩阵。Word2Vec和Glove生成的M维特征的权重矩阵分别如式(1)和式(2)所示。

$$\mathrm{DW}=\{w_1,w_2,\cdots,w_M\} \tag{1}$$

$$\mathrm{DG}=\{g_1,g_2,\cdots,g_M\} \tag{2}$$

其中，w_i和g_i权重比例对应于第i个特征。**DW**与**DG**并分别归一化以生成**DW'**和**DG'**，归一化公

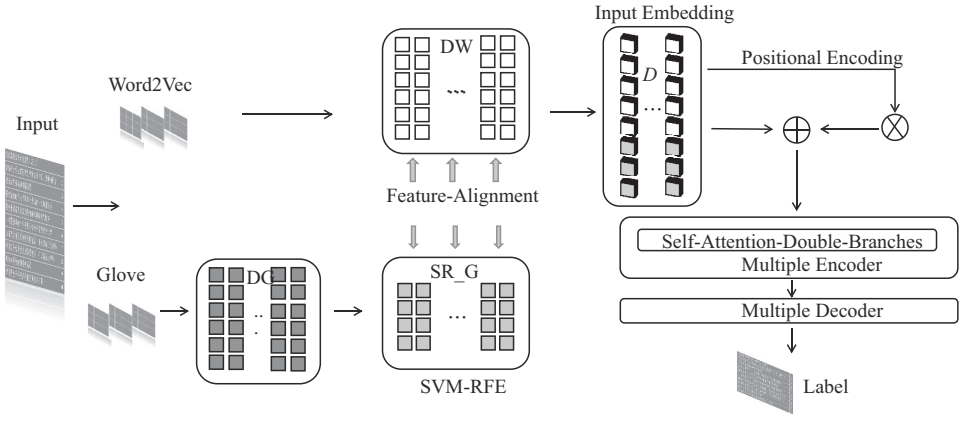

图1　模型整体架构

式为

$$w_i' = \frac{w_i - w_{\min}}{w_{\max} - w_{\min}} \tag{3}$$

其中，w_i'为归一化后的数值，$w_{\min}$为词嵌入最小值，$w_{\max}$为词嵌入最大值。

2.2　SVM-RFE提取特征向量

SVM为有监督的分类算法，算法核心是求最优超平面$(w\cdot x+\mathrm{b}=0)$进而将目标群体进行划分。在处理非线性高维度分类问题中，SVM使用核函数将维度提升，一直提升至线性可分。对于一个二分类问题$y_i=\pm 1$，$(x_i,y_i)(x_i\in R^p,y_i\in[-1,1],i=1,2,\cdots,n)$表示由$n$个样本和$p$个特征组成的训练集，并且$\mathrm{y_i}$是训练样本$\mathrm{x_i}$的标签，$\xi$为松弛变量。其目标函数为

$$J=\frac{1}{2}w^Tw+\mathrm{C}\sum_{i=1}^{n}\xi_i \tag{4}$$

$$y_i(w^Tx_i+\mathrm{b})\geqslant 1-\xi_i,\xi_i\geqslant 0,i=1,2,\cdots,n \tag{5}$$

为了使超平面的划分更加准确，引入拉格朗日函数，如式(6)所示。

$$L(w,b,\xi,\alpha,\gamma)=\frac{1}{2}w^2+\mathrm{C}\sum_{i=1}^{n}\xi_i-\sum_{i=1}^{n}\alpha_i[y_i(wx_i+\mathrm{b})-1+\xi_i]-\sum_{i=1}^{n}\xi_i\gamma_i \tag{6}$$

其中，$\alpha_i\geqslant 0,\gamma_i\geqslant 0$。预测未知样本$\boldsymbol{x}$的分类结果通过决策函数$f(x)=y(w^Tx+\mathrm{b})$来决定，其中$\boldsymbol{w}=[w_1,w_2,\cdots,w_p]$为特征权重向量。当删除特征$\boldsymbol{W}$时，与之对应的特征变化为$W_i=\boldsymbol{W}+w_i^2$。因此，当找到最小值$w^2$时，删除该特征便可以同时实现优化特征和造成最小的影响。将Glove算法生成的词向量$\mathbf{DG}=\{g_1,g_2,\cdots,g_M\}$送入SVM-RFE算法中，经过特征排序后，取前$L$维生成词特征向量$\mathbf{SR_G}$。将$\mathbf{SR_G}$与$\boldsymbol{W}$词向量融合拼接，送入后续模型进行分类。

2.3　多分支注意力模型

在Transformer模型中，位置编码应用于编码器与解码器前的输入中，使用正弦和余弦函数计算。Transformer模型的自我注意层可以被视为一种点积注意机制，可以将其描述为将查询$\boldsymbol{Q}$和一组键值对$\boldsymbol{K}$和$\boldsymbol{V}$映射到一个实数，其中$\boldsymbol{Q}$、$\boldsymbol{K}$和$\boldsymbol{V}$都是向量。具体来说，给定一个长度为l、维数为d的向量序列$\boldsymbol{H}\in R^{l\times d}$作为输入序列，将自注意力矩阵投影到三个不同的矩阵中：查询矩阵$\boldsymbol{Q}$、关键矩阵$\boldsymbol{K}$和值矩阵$\boldsymbol{V}$。然后，应用点积注意力来获得输出表示如下

$$\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}=\boldsymbol{HW}^Q,\boldsymbol{HW}^K,\boldsymbol{HW}^V \tag{7}$$

$$\mathrm{Attention}=(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{soft\,max}\left(\frac{\boldsymbol{QK}^T}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{8}$$

其中，$\boldsymbol{HW}^Q,\boldsymbol{HW}^K,\boldsymbol{HW}^V$是需要学习的模型参数，$\left(\frac{\boldsymbol{QK}^T}{\sqrt{d_k}}\right)$是自注意力分数。本文基于Transformer模型修改了注意力模型的计算机制，旨在通过计算神经网络模型的不同深度来提取更有效的信息。双通道注意力模型如图2所示，词向量$\mathbf{WordVector_d}^{i1},\mathbf{WordVector_d}^{i2}$，乘以两组生成的矩阵$(\boldsymbol{W}^{q1},\boldsymbol{W}^{k1},\boldsymbol{W}^{v1})$、$(\boldsymbol{W}^{q2},\boldsymbol{W}^{k2},\boldsymbol{W}^{v2})$经过矩阵融合运算，生成$\mathbf{WordVector_b}^{i1}$。

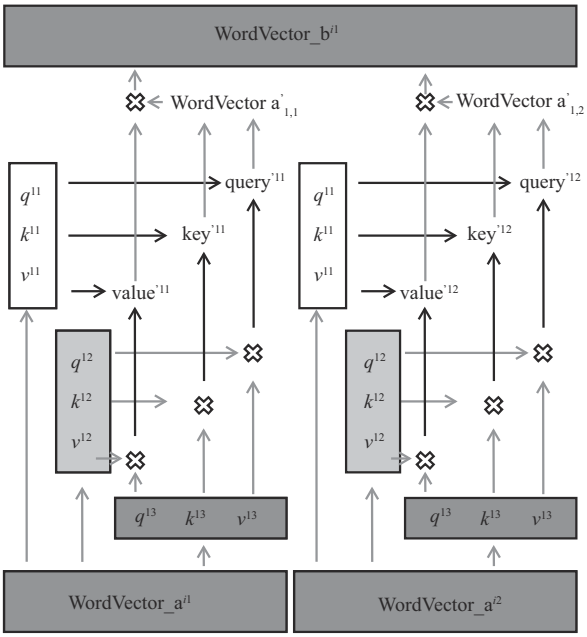

图2　双通道注意力模型

为了提高自我注意模型的性能，Transformer采用了多头自注意机制，该机制结合了从同时训练的多个子空间中学习到的信息，同时保持模型参数不变，如式(9)与式(10)所示。这种方法旨在提取尽可能多的信息，同时减少模型过拟合。

$$\mathbf{MHA}=[\mathrm{head}_1,\mathrm{head}_2,\cdots,\mathrm{head}_H]\boldsymbol{W}^O \tag{9}$$

$$\mathrm{head}=\mathrm{Attention}(\boldsymbol{HW}_i^Q,\boldsymbol{HW}_i^K,\boldsymbol{HW}_i^V) \tag{10}$$

其中，参数$\boldsymbol{W}^O,\boldsymbol{W}_i^Q,\boldsymbol{W}_i^K,\boldsymbol{W}_i^V$是需要学习的参数之一。多头注意机制通过计算$H$个独立注意头上的注意来扩展传统的注意机制。这使得模型能够专注于来自不同位置的不同子空间的信息，从而避免陷入局部最优。

3 实验与分析

3.1 山东大学邮件数据集

本文独立收集了两千余封大学电子邮件记录，在处理数据后，将电子邮件信息分为三个主要主题。这些主题又分为多个子主题：资讯类主题包括15类，注册类包括5类，其他情况包括4类。这些主题涵盖的主要主题包括招生、转学、录取和招生制度。高校电子邮件的主要内容与分类如表1所示。

3.2 公共数据集

该算法使用的公共数据集是IMDB数据集和THUCNews数据集。

1) IMDB数据集。IMDB数据集是文本分类中备受推崇的资源，是文本分类任务的基准。这是一个广泛的在线数据库，提供有关电影、电视节目、演员、导演、编剧和制片人的信息。该数据集包括50 000条IMDB评论，专门用于情绪分析。

2) THUCNews数据集。THUCNews是清华大学和新浪新闻合作开发的一个综合数据集。它由来自新浪新闻RSS订阅频道的过滤和筛选的历史数据组成，涵盖2005年至2011年。THUCNews包含74万个UTF-8纯文本格式的新闻文档。

3.3 邮件分析实验

为了确认双通道注意机制及SVM-RFE的效果，本文使用不同的策略进行了实验，概述如下。

1) Transformer：训练和测试Transformer模型。

2) Transformer-TBAM：用双通道注意力机制改进Transformer模型。

3) Transformer-SR（SVM-RFE）：通SVM-RFE改进Transformer模型。

4) SVM-RFE-Transformer-TBAM：通过SVM-RFE和双通道注意力机制改进Transformer模型。

在计算了本文中的文本分类算法后，在公共数据集上获得的最终结果如表2所示。

表2 公共数据集实验结果对照表

算法名称	IMDB准确率	THUCNews准确率
Transformer	0.906 2	0.890 5
TextCNN[10]	0.910 3	0.910 6
TextRCNN[11]	0.912 1	0.923 8
Transformer+TB	0.910 7	0.895 1
Transformer+SR	0.920 1	0.902 5
SVM-RFE-Transformer-TBAM	0.925 7	0.927 2

表3显示了模型在测试集上的准确性。可以看出，使用经过SVM-RFE筛选后的特征向量的双通道注意力机制可以提高基于Transformer的文本分类模型在数据集上的性能，并且显著优于仅使用注意力机制的模型的性能。这进一步证实了所提出方法的有效性。在大学电子邮件数据集上的测试性能为94.67%。

表3 大学邮件数据集实验结果比较

算法名称	准确率
Transformer	0.934 7
TextCNN	0.936 1
TextRNN	0.938 5
Transformer-TB	0.941 2
Transformer-SR	0.938 2
SVM-RFE-Transformer-TBAM	0.946 7

4 结束语

本文提出了一种结合SVM-RFE与Transformer-TBAM用于大学电子邮件文本的情感分析的算法，利用SVM-RFE技术筛选出词向量中的关键特征，并借鉴注意机制和卷积神经网络的原理，对Transformer模型中的自注意机制进行了优化，旨在提升文本情感分类的准确度和泛化性能。本文算法采用

表1 高校电子邮件的主要内容与分类

项目类别	邮件主要内容				
咨询类	本科生招生政策	招生专业及目录	研究生招生名额	初试成绩	录取通知书
	复试安排	放弃资格	入住须知	博士生招生政策	审核评估办法
	报名登记表	博士生导师	保密方案	大纲调整	建议录取
注册问题	系统无法登录	取消报名	无法提交材料	退款费用	系统开放时间
其他类	举报	广告	违规行为	投诉	

独特的双通道架构，能够深入挖掘文本中的深层信息，进而增强了模型处理单词间依赖关系的能力。同时，它还融合了 Word2Vec 和 Glove 模型来训练单词嵌入，从而从文本中提取出更丰富的信息。

为了评估所提方法的有效性，本文在 2 个公开的数据集以及一个自行构建的大学电子邮件数据集上进行了实验。实验结果显示，该方法在情感分类任务中的表现优于传统的 Transformer 模型，展现出了卓越的性能和泛化能力。该算法在大学电子邮件文本分类任务中的准确率达到了 94.67%，相较于传统文本分类算法提高了 1.2%。此外，在其他公开数据集上，该方法也表现出了强大的泛化能力。多分支注意力机制有效地挖掘了更深层次的信息，并增强了从多个子空间中收集信息的能力，从而进一步提升了文本分类的性能。

参考文献：

[1] 孟祥福,石皓源 . 基于 Transformer 模型的时序数据预测方法综述[J/OL]. 计算机科学与探索, (2024-07-30)[2024-10-22].

[2] 闫芳序, 王剑辉 . 基于 SVM 和 Word2Vec 的微博评论情感识别模型[J]. 现代计算机, 2024, 30(10): 60-64.
YAN F X, WANG J H. Sentiment recognition model of weibo comments based on SVM and Word2Vec[J]. Modern Computer, 2024, 30 (10): 60-64.

[3] PENNINGTON J, SOCHER R, MANNING C. Glove: global vectors for word representation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL Press, 2014: 1532-1543.

[4] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86 (11): 2278-2324.

[5] ZAREMBA W , SUTSKEVER I , VINYALS O .Recurrent Neural Network Regularization[J]. arXiv Preprint, arXiv: 1409.2329, 2014.

[6] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9: 1735-1780.

[7] UMER M, IMTIAZ Z, AHMAD M, et al. Impact of convolutional neural network and FastText embedding on text classification[J]. Multimedia Tools and Applications, 2023, 82(4): 5569-5585.

[8] BETUL POLAT S, CANKURT S. Fake news classification using BLSTM with glove embedding[C]//Proceedings of the 2023 17th International Conference on Electronics Computer and Computation (ICECCO). Piscataway: IEEE Press, 2023: 1-5.

[9] LI X L, ZHANG Y Y, JIN J, et al. A model of integrating convolution and BiGRU dual-channel mechanism for Chinese medical text classifications[J]. PLoS One, 2023, 18(3): 1-20.

[10] YOON K. Convolutional neural networks for sentence classification [C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL Press, 1746-1751.

[11] LAI S W, XU L H, LIU K, et al. Recurrent convolutional neural networks for text classification[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2015, 29(1): 2267-2273.

[作者简介]

李振（1995– ），男，山东泰安人，山东大学工程师，主要研究方向为教育信息化、大数据分析。

李智超（1989– ），男，山东滨州人，山东大学工程师，主要研究方向为高校信息化、软件工程、人工智能。

陈琳（1983– ），男，山东济南人，博士，山东大学高级工程师，主要研究方向为教育信息化、人工智能。