面向不平衡短文本情感多分类的三阶语义图数据增广方法

颜学明[1)] 黄 翰[2),3),4)] 金耀初[5)] 钟 国[1)] 郝志峰[6)]

[1)](广东外语外贸大学信息科学与技术学院 广州 510006)

[2)](华南理工大学软件学院 广州 510006)

[3)](大数据与智能机器人教育部重点实验室 广州 510006)

[4)](广东省大模型与生成式人工智能技术工程中心 广州 510006)

[5)](西湖大学工学院 杭州 310030)

[6)](汕头大学数学与计算机学院 广东 汕头 515063)

摘 要 文本增广技术可以有效提升不平衡情感分类任务的性能.若文本增广过程中生成的少数类短文本数据未能体现完整的情感语义特征,则可能会导致不同类别之间的情感重叠问题出现.为了充分学习和理解少数类别的情感特征,本文提出一种面向不平衡文本情感多分类的三阶语义图数据增广方法,首先采用三阶语义图在多个词之间建立复杂的关系语义模型,用于表示多种可能的短文本局部情感语义和词节点依赖关系,然后提出了基于三阶语义图数据增广方法以平衡多分类文本的情感类别分布,从而有效实现不平衡短文本的情感分类.与传统的文本增广方法相比,在印尼语不平衡数据集上,本文提出的方法在少数类评价指标F_1-measure和F_2-measure上分别提升了5.75%和9.65%,在平衡情感识别能力指标G-means值上提升了2.91%;在马来语不平衡数据集上,本文提出的方法在少数类评价指标F_1-measure和F_3-measure上也分别提升了2.45%和4.81%,在平衡情感识别能力指标*G-means*值上提升了1.24%.此外,与传统的机器学习方法、深度网络模型等情感分类模型以及传统的短文本增广过采样模型相比,本文提出的方法在公开的印尼语、马来语、英语以及中文四个不平衡短文本数据集上都获得了最高的准确率*Accuracy*值.以上实验结果表明,融合不同模体的三阶语义图结构信息不仅可以有效表达文本中的局部情感语义以及词节点之间的依赖关系,还可以有效降低短文本数据增广过采样过程中引入新噪声的风险,并提升不平衡短文本的多分类性能.

关键词 三阶语义图;文本增广;平衡策略;短文本情感分类;模体

中图法分类号 TP18 **DOI号** 10.11897/SP.J.1016.2024.02742

A Short Text Augmentation Approach Based on Three-Order Semantic Graphs for Imbalanced Sentiment Multiclassification

YAN Xueming[1)] HUANG Han[2),3),4)] JIN Yao-Chu[5)] ZHONG Guo[1)] HAO Zhi-Feng[6)]

[1)](*School of Information Science and Technology, Guangdong University of Foreign Studies, Guangzhou* 510006)

[2)](*School of Software Engineering, South China University of Technology, Guangzhou* 510006)

[3)](*Key Laboratory of Big Data and Intelligent Robot , MOE of China, Guangzhou* 510006)

[4)](*Guangdong Engineering Center for Large Model and Generative Artificial Intelligence Technology, Guangzhou* 510006)

收稿日期:2023-06-01;在线发布日期:2024-09-20.本课题得到国家自然科学基金重点项目(62136003)、国家自然科学基金项目(62276103,62476163)、国家自然科学基金合作创新研究团队项目(W2441019)、广东省基础与应用基础研究基金(2023B1515120020)和广东省普通高校创新团队项目(2023KCXTD002)资助.**颜学明**,博士,副教授,中国计算机学会(CCF)高级会员,主要研究领域为自然语言处理、机器学习以及图优化.E-mail:yanxm@gdufs.edu.cn.**黄 翰**(通信作者),博士,教授,博士生导师,中国计算机学会(CCF)杰出会员,主要研究领域为微计算理论与方法、智能化软件工程、数据智能工程.E-mail:hhan@scut.edu.cn.**金耀初**,博士,讲席教授,博士生导师,长江学者、IEEE Fellow,中国计算机学会(CCF)会员,主要研究领域为数据驱动的复杂系统进化优化、进化多目标机器学习、联邦学习与安全机器学习以及演化发育系统与形态发育机器人学.**钟 国**,博士,副教授,中国计算机学会(CCF)会员,主要研究领域为数据挖掘、机器学习.**郝志峰**,博士,教授,博士生导师,中国计算机学会(CCF)会员,主要研究领域为机器学习、计算智能、代数学与组合优化.

[5](*School of Engineering, Westlake University, Hangzhou* 310030)
[6](*College of Mathematics and Computer Science, Shantou University, Shantou, Guangdong* 515063)

Abstract Text augmentation techniques have been widely recognized for their ability to significantly enhance the performance of sentiment classification tasks, particularly when dealing with imbalanced datasets. However, when generating short text data for the minority class during text augmentation, it can result in overlapping emotions across different categories if the generated data fails to capture the complete semantic features of sentiment. To better understand and represent the emotional features of minority classes, this study introduces a third-order semantic graph data augmentation method specifically designed for imbalanced text sentiment multi-classification. The proposed method is based on the construction of a third-order semantic graph that models complex relationships between multiple words within short texts. The proposed method allows for the representation of a wide range of local sentiment semantics and is able to capture the dependencies between word nodes, offering a more nuanced understanding of emotional context in minority classes. By leveraging this intricate relational model, the third-order semantic graph enables a more comprehensive representation of sentiment, ensuring that the emotional characteristics of minority classes are more accurately reflected in the generated data. Once the third-order semantic graph is constructed, a data augmentation method based on this graph is applied to balance the distribution of sentiment categories in multi-class text datasets. This approach is designated to address the shortcomings of traditional text augmentation methods that often introduce noise and fail to adequately represent minority class sentiments by ensuring that the generated text data can capture the essential emotional features of the minority class, thus leading to improved classification performance across imbalanced datasets. Compared with traditional text augmentation methods, the proposed method in this paper can improve the minority evaluation indicators F_1-measure and F_2-measure by 5.75% and 9.65%, respectively, and the *G-means* value of balanced emotion recognition ability by 2.91% on the unbalanced Indonesian dataset. On the unbalanced Malay dataset, the proposed method also increases the minority evaluation indicators F_1-measure and F_3-measure by 2.45% and 4.81%, respectively, and the *G-means* value of balanced emotion recognition ability by 1.24%. In addition, compared with existing sentiment classification models based on short text augmentation oversampling models, traditional machine learning methods and deep network models, the proposed method achieves the highest *Accuracy* values on the publicly available imbalanced short text datasets in Indonesian, Malay, English, and Chinese. The experimental results also demonstrate that the proposed method provides a comprehensive and effective solution to the challenges posed by imbalanced sentiment classification. By integrating third-order semantic graph structures across different modalities, the proposed method effectively captures local emotional semantics and word dependencies. This not only improves the representation of minority class emotions but also significantly reduces the risk of introducing noise during data augmentation. In addition, traditional oversampling methods often introduce errors that can degrade classification performance, whereas the proposed method avoids these pitfalls by leveraging the detailed relational structure of the semantic graph. As a result, it can also achieve better multi-class classification performance on imbalanced short text sentiment classification tasks.

Keywords third-order semantic graphs; text augmentation; balancing strategies; short text sentiment classification; motif

1 引 言

随着互联网技术的普及，社交媒体在人们日常生活中发挥着越来越重要的作用．越来越多的用户在Twitter、Facebook和微博等社交平台上表达他们的思想观点和发表评论．社交媒体上产生的短文本评论数据是用户根据客观事物的真实感受以及个人经验形成的意见与观点，具有一定的主观性[1]。此外，大规模短文本数据的情感类别还存在一定的分布不平衡性．例如，当用户在社交媒体上分享事实或信息时，通常不会表达强烈的情感，因此中性类评论的数据往往比积极类评论和消极类评论都要多[2-3]．这种类别不平衡问题给短文本情感分类带来了巨大的挑战．

在求解不平衡短文本情感分类问题时，传统的机器学习方法大都是侧重多数类而忽略少数类[4]．然而在大多数现实情况中，正确识别少数类样本的情感信息往往更具有价值．因此，少数类短文本数据的情感信息才是不平衡情感分析需要重点关注的对象[5]．有学者尝试通过改进传统的机器学习算法或者提出新算法以求解不平衡短文本情感分类问题，但在短文本数据不平衡的情况下，传统算法提升的分类效果比较有限[6]．此外，还有学者提出，平衡数据的策略可以提升不平衡短文本情感分类方法的性能，主要包括对多数类欠采样或者对少数类过采样这两种类别平衡策略．欠采样平衡策略指通过从多数类样本中删除一些样本来平衡不同类样本的样本总数[7]．然而，不恰当的裁剪可能会删除一些潜在有用数据，使分类器丢掉有关多数类的重要性，导致不同类边界重叠的程度更加严重．此外，由于欠采样策略与分类评价过程是相对独立的，欠采样平衡数据的方法可能会导致欠采样与分类训练之间信息不对称．

与欠采样平衡策略不同，过采样的平衡策略是在少数类样本中增加一些样本，从而平衡不同类别的样本总数，提升分类方法的性能．例如，基于文本增广过采样的情感分类平衡策略[8]不仅可以平衡训练集的类分布，还可以修改噪声数据．从目前的研究来看，基于文本增广的过采样平衡策略在不平衡短文本数据的处理上更具优势．然而，若短文本数据不能用精确的词向量模型来表示，则生成的少数类短文本数据可能会存在不完整的情感语义[9]．如图1所示，假设有一份产品评论的数据集，其中包含了用户对数码产品的评论以及相应的情感标签，包括积极情感(Positive)、消极情感(Negative)和中性情感(Neutral)．假设在数据集中积极情感和消极情感为少数类，因此需要增广生成相应的文本数据．在增广评论1中，原始评论中的“love”被改为“adore”，虽然它们在情感上相似，但它们的语义可能略有不同．因此，无法确保增广后的评论仍然包含完整的语义特征．这种语义差异可能导致情感分类模型在处理增广后的评论时性能下降．同时，过采样平衡策略产生的新数据质量可以直接影响不平衡短文本的分类方法性能，但可能存在类别边界重叠的缺点[10]．如图1所示，增广评论2中存在情感类别重叠问题，评论中先提到价格低廉的中性情感，再提到画质好的积极情感．这种情感重叠问题使情感分析任务变得更加复杂，因为分类模型难以明确区分这些重叠的情感类别．

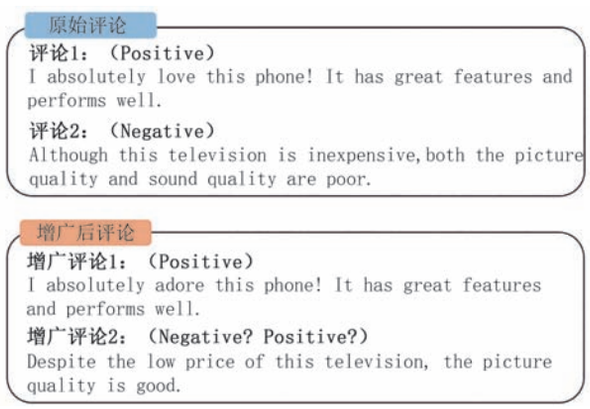


图1 基于过采样文本增广前后的评论实例

鉴于此，本文提出一种三阶语义图数据增广方法(A Three-order Semantic Graphs Augmentation Approach，TSGA)，以提升不平衡短文本的情感预测与多分类性能．该方法先对短文本进行三阶语义图表示，然后设计基于三阶语义图情感语义的数据增广方法，用于生成能够包含多种局部情感语义信息的少数类短文本数据，实现不平衡短文本数据的情感分类．本文的主要贡献如下：

(1)提出了一种基于三阶语义图的短文本情感语义表示模型．该方法以三阶模体作为词节点局部语义关系的研究对象，并在多个词之间建立复杂的情感关系语义模型，使文本增广过程中所生成的新样本与原始样本具有一定的情感语义相关性，减少冗余样本的生成．

(2)提出了一种基于不平衡短文本多分类问题的三阶语义图数据增广方法．在该方法中，基于三

阶语义图的文本增广算法可以在一定程度上避免类别间的数据重叠问题，进而平衡数据集的分布并增加多分类训练数据的多样性，提升不平衡短文本情感多分类任务的分类性能.

本文第2节介绍短文本情感分析与基于文本增广方法的平衡策略相关工作. 第3节重点介绍本文所提出的基于三阶语义图数据增广的不平衡短文本多分类方法. 第4节是本文的对比实验以及对实验结果的详细阐述与分析. 最后一节是对本文工作的总结和未来展望.

2　相关工作

本节先介绍短文本情感分类的相关工作，然后介绍基于文本增广的平衡策略的不平衡短文本情感分类相关工作.

2.1　短文本情感分类

目前，短文本情感分类的方法主要包括基于词典的情感分类方法、基于机器学习的情感分类方法以及基于深度学习的情感分类方法.

基于词典的学习方法利用已经构建好的情感词典来判断短文本的情感类别[11]. 例如，Chekima等人[12]构建了一个马来语情感词典，用于马来语社交平台的评论进行情感分析. Lailiyah等人[13]基于已有的印尼情感词典对印尼民众评论进行了情感分类，该方法对具有嘈杂信息的印尼语社交评论数据的分类准确率并不高. 虽然基于词典的情感分类方法处理速度快，但过于依赖情感词典的质量，具有一定的局限性.

对基于机器学习的方法而言，短文本的情感分类问题是一个特殊的文本分类问题. 由于不受词典规模和更新的限制，基于机器学习的方法可自动提取文本特征. 张仰森等人[14]从统计学方法与机器学习相结合的角度出发，提出一种级联式微博情感分类模型来提高情感分析的准确度. Tanasanee等人[15]构造了有效的多核函数，以提高支持向量机的情感分类性能. Fiarni等人[16]采用了基于规则和朴素贝叶斯算法的方法对印尼在线运输服务评论进行情感分类. Sadanandan等人[17]采用知识库和多个情感分类器集成的方法提升马来语短文本情感分类性能. Boiy等人[18]从多语言中提取顾客评论的情感信息，并使用多种机器学习集成等方法对其进行分类. 虽然这些基于机器学习的方法能够对某种特征进行建模，但其分类性能还是比较依赖短文本情感数据特征的表示与建模[19].

基于深度学习的方法通过自动学习并获取短文本数据的丰富语义信息，取得了比传统机器学习方法更好的情感分类效果. 何炎祥等人[20]提出一种新颖的情感分类模型，可以有效增强卷积神经网络捕捉文本情感特征信息的能力. Wang等人[21]在不需要逐层预训练的情况下提取时序化的短文本情感特征，然后采用词嵌入模型学习短文本中的上下文信息，最后使用LSTM算法来解决社交媒体中的短文本情感分类问题. 然而，这些序列化编码网络并不能直接从短文本的句法结构、语义关系图等信息中提取深度学习特征[22]，因此有学者[23-25]通过构建合适的图网络结构表示模型来提取短文本情感分类中的关键信息. Yao等人[23]在词共现和文档-词关系的基础上为语料库构建一个文本图结构信息，并提出基于短文本的图卷积神经网络的分类方法. Hu等人[24]提出一种基于异构图注意力网络的方法以对短文本进行分类，其中的图注意力机制可以学习当前节点到不同相邻节点以及不同类型节点的重要情感语义信息. Hajibabaee等人[25]提出基于有向graphSAGE和word2vec的图算法以预测短文本类别. 虽然这些图网络结构的深度学习方法可以从不同的角度学习到丰富的短文本情感特征，但在语言资源不足或者缺失的情况下，提升短文本情感多分类的性能仍然具有一定的挑战性. 这主要是因为不准确的词向量表示可能难以获取短文本中词节点之间潜在的完整情感语义.

2.2　基于文本增广的平衡策略

在不同领域的社交媒体上，短文本评论信息通常具有一定的情感不平衡性，给短文本情感分析带来了巨大的挑战. 在短文本数据集规模普遍不大的情况下，通过不平衡学习增加少数类样本的过采样方法，在求解类别数据之间存在分布不均衡的问题上是比较有效的[26]. 若过采样方法随机复制现有的少数类样本，虽可以平衡数据，但训练出来的模型会存在一定的过拟合问题. 新样本数据的生成过程容易引入新的冗余或者无效噪声，甚至导致类别间的重叠程度加大[27]. 因此，过采样方法中的少数类短文本数据生成策略是非常重要的.

为了生成有效的少数类样本，近年来不少学者陆续提出了一些基于数据增广的方法[8]，主要包括标签无关的增广方法和标签相关的增广方法[27]. 标签无关的增广方法是指只需基于无标签的训练数据并按照规则来实现数据增广，不需要提供数据标签、任务需求等信息. 例如，Liu等人[28]在邮件文本分类

任务中结合了基于近义词表和基于词向量的单词替换方法，但该方法仅解决了近义词表只能应用于特定范围内单词的问题. 标签相关的增广方法是指利用标签信息并按照任务需求来对文本进行增广，还需要考虑增广数据的标签与原数据标签是否发生变化. 例如，Cheng等人[29]先按照单词替换的方法将两条短文本样例生成对抗样例，之后提出不受数据标签限制且可跨标签增广的混合方法. Yan等人[30]使用随机排序的方法对法律文书进行句子级别的操作，由于句子独立地包含了相对完整的语义，且文书中句子的顺序对原始文本的含义影响不大，因此可以通过将句子打乱顺序并进行随机排列来得到增广文本. 在具有特殊数据特征的任务中，标签相关的增广方法则需要根据需求选择合适的数据增广方法，生成新的少数类样本以扩大少数类训练数据，从而平衡数据集[27]. 在标签相关的文本增广方法中，采用词图结构对单词的词频信息和上下文信息进行表示，可以在一定程度上有效获取词节点之间的局部潜在语义关系[31]. 然而，由于词图的不规则结构，当前的图数据增广策略往往需要对整个图数据集进行改动，因此很难直接面向不平衡情感多分类问题来生成小数类文本数据[32].

受Chen等人[31]的启发，本文先对短文本的局部情感语义和词依赖关系进行三阶语义图集表示. 为有效实现不平衡短文本数据的情感类别平衡分布，本文提出一种基于三阶语义图文本增广方法来平衡数据并对多分类模型进行训练，从而有效提升不平衡短文本的情感分类性能.

3 三阶语义图模型及其增广方法

本节主要介绍一种基于三阶语义图数据增广的不平衡短文本情感多分类方法. TSGA模型的基本框图如图2所示. 首先，对不同三阶模体下短文本的不同情感语义关系进行建模，在词图表示基础上采用三阶语义图模型对短文本进行表示. 其次，为了平衡不同情感类别的短文本数据分布，提出基于三阶语义图数据的增广方法以生成新的少数类短文本数据. 最后，对不平衡短文本的三阶语义图结构信息进行多分类模型训练，并得到最终的情感类别. 与传统的文本增广方法不同，本文提出TSGA算法可以在多个三阶模体下获取短文本词图的局部情感语义和词依赖关系图，并从一定程度上降低数据增广过程中引入新噪声以及类别重叠的风险. 论文中提及的主要符号及说明如表1所示.

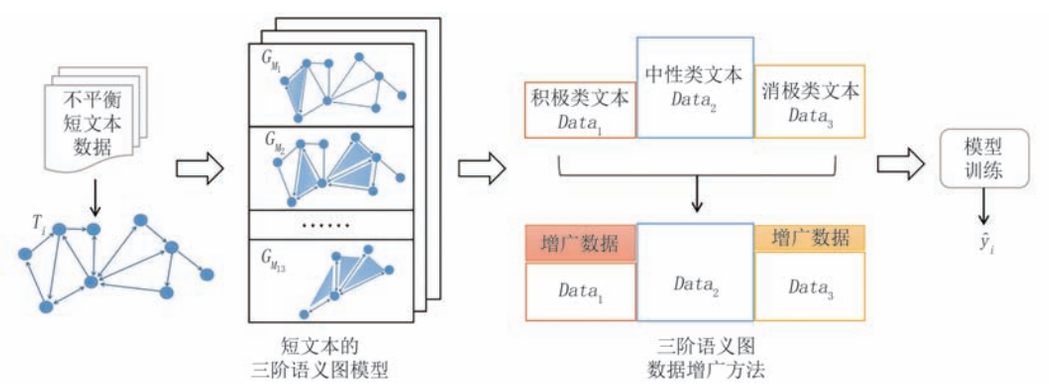


图2 基于三阶语义图数据增广的不平衡短文本情感分类

表1 符号说明

符 号	备 注
T	短文本数据集
N	数据集中的样本数
$G=(V,E)$	词图的二元组表示
G_M	词图G的三阶语义图结构表示
X	词节点的初始语言特征集
Y	情感类别集
r_j	不平衡率
ρ	不平衡度参数

3.1 任务定义

设给定的短文本数据集$T=\{T_i|1\leqslant i\leqslant N\}$，$Y=\text{Label}(T)$是$T$对应的情感类别标签集. 例如，$Y=\{1,0,-1\}$指的是积极、中性和消极这三类情感类别标签集. 在一些短文本twitter数据集中，不同情感标签的类别分布会存在一定程度的不平衡性. 例如，在twitter评论数据集中[2]，大部分文本的观点都是中性态度. 此外，一些语言短文本中还可能混有少量其他语言的词[33]. 以英语短文本“I was

delighted to try the sate and nasi goreng”为例，该句子含有3个印尼语单词“sate”“nasi”和“goreng”. 语言资源不足以及短文本混有其他语言特征信息等问题都可能导致短文本数据不能生成准确的词向量表示，这进一步增加了不平衡短文本情感分类问题的难度. 本文通过分析和挖掘不平衡短文本 T_i 的三阶语义图结构情感语义，自动将不具有精确词向量表示的短文本分类到多个先验的情感标签类别中.

3.2　基于短文本词图表示的三阶语义图模型

词图(Graph-of-words)作为短文本的一种表示形式，在自然语言翻译、文本挖掘等方面也取得了不错的效果[34]. 在解决某些领域的应用问题时，词图可以有效捕获词与词之间的某种潜在关系，并在一定程度上减少对语言特征的部分依赖性. 本文在短文本词图的基础上将模体作为词节点局部语义关系的研究对象，从三阶的角度提取短文本词图结构的有用知识和信息.

定义1.　短文本的词图为二元组 $G=(V,E)$. $V=\{v_i|1\leqslant i\leqslant n\}$ 表示词节点集，v_i 为词节点，n 为短文本中含有的词节点数量. E 表示词节点之间的边集，$E=\{e_{ij}|\ 1\leqslant i\leqslant n,\ 1\leqslant j\leqslant n,\ i\neq j\}$ 表示有向边集.

有向边 e_{ij} 表示词 v_i 与 v_j 之间的词序关联性. 在词图中，词节点之间存在某些特定的子图结构，子图结构对应的局部情感语义具有一定的传递性与稳定性[35]. 当词 v_i 与 v_j 同时出现在 w 大小的窗口时，它们在窗口 w 内存在某种潜在语义关系[34]. w 的值越大，则捕获词之间的依赖关系越多，词图的复杂性也会越大[36]. 以英语文本“I love the flowers as love you”为例，该句子包括“I”“love”“the”“flowers”“as”以及“you”6个词节点. 当窗口大小分别为2和3时，词图表示分别如图3(a)和图3(b)所示. 在图3(a)的词图中，除了可以表示词节点的信息之外，还能通过边(二阶模体)表示词节点之间的上下文语义. 图3(b)显示的是窗口 $w=3$ 的词图结构，可以看出除了可以表示图3(a)蕴含的上下文语义，还可以有更复杂的上下文语义.

定义2.　模体是指一个有 $n(n>1)$ 个节点的有向连通图. 模体是一种特殊的子图，由节点和节点之间的边组成，其节点和边的组合和排列方式通常可以承载更多的结构语义信息，而不仅仅是单独的节点或边所能表达的信息[37].

由于模体通常是基于文本的上下文或相邻节点

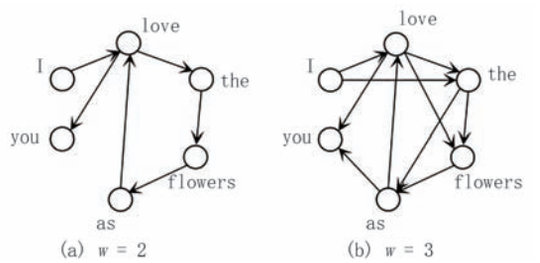


图3　不同窗口 w 下短文本词图结构

来定义的，因此它们可以反映出文本中词汇之间的高阶关联性以及特定的语义关系，从而用于识别文本中不同形式的情感特征. 特别地，具有三个或更多节点(即 $n\geqslant 3$)的模体可以捕捉高阶图结构. 其中，对于给定节点，其模体具有 $n>3$ 的节点(即不仅包括与其相邻节点相邻的边，还包括其相邻节点之间的边)所捕获的高阶结构，可以被多个包含3个节点的模体类似地捕获[38]. 因此，给定3个节点的模体具有足够的能力来表示文本的情感语义结构. 图4展示了与窗口大小为3的词图结构对应的13种三阶模体的存在形式.

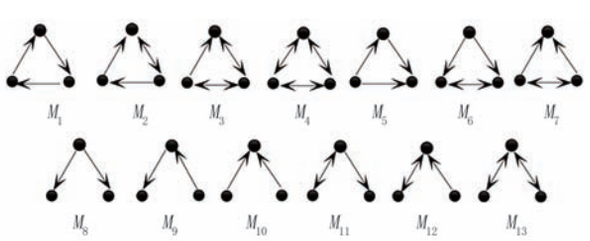


图4　三阶模体的所有存在形式

定义3.　词图 $G=(V,E)$ 的三阶语义图的集合为 $G_M=\{G^{(k)}|1\leqslant k\leqslant 13\}$，其中 $G^{(k)}=(V,E,W_{M_k},X)$ 为基于模体 M_k 的三阶语义图. $W_{M_k}=\{w_{ij}^{(k)}\}_{n\times n}$ 为基于模体 M_k 的邻接矩阵，$w_{ij}^{(k)}$ 为边 e_{ij} 在选定模体 M_k 中出现的次数. 特别当 $W_{M_k}=0$ 时，我们定义 $G^{(k)}=\varnothing$. $X=\{x_i|1\leqslant i\leqslant n\}$ 为词图节点的初始文本语言特征集，x_i 指词节点 v_i 的初始文本特征，n 为三阶语义图中词节点的数量.

通过对不同三阶模体下的局部子图的情感语义以及词依赖关系进行更好的重构与学习，用模体代替短文本词图的单个词节点，可以得到短文本词图的三阶语义图结构表示. 图5展示了图3(b)中词图的三阶语义图结构表示过程. 首先，在词图的基础上，我们可以得到三阶模体下的词图邻接矩阵. 在基于三阶模体 M_5 的邻接矩阵 W_{M_5} 中，$w_{23}^5=2$ 表示边 e_{23} 在选定的 M_5 模体中出现过2次. 若某条边在选定模体中出现的次数越多，则表示基于该边的词

节点越有可能在选定模体中可以更准确表示局部情感语义.

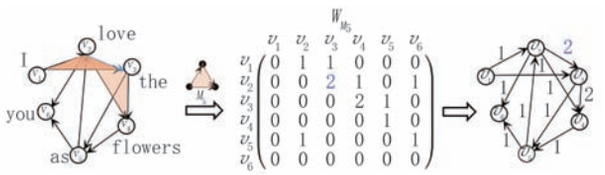


图5 基于三阶模体M_5的三阶语义图结构表示

为了提取更准确的情感语义信息,我们采用Seg-Bert预训练模型[39]学习得到不同模体下的三阶语义图结构情感特征.通过预训练过程,原始特征x_i经过Seg-Bert编码层可以得到的隐层表示为z_i,再经过一层FC映射后重建原始特征得到$\hat{x}_i = FC(z_i)$.为确保重建后的特征$\hat{x}_i$可以尽可能捕捉到不同模体的子图结构情感信息,三阶语义图结构信息重构的损失函数$\mathcal{L}$为

$$\mathcal{L}=\frac{1}{|\mathcal{V}_{M_k}|^2}\|\bar{W}_{M_k}-\hat{S}\|_2 \tag{1}$$

其中$\mathcal{V}_{M_k}$为基于模体M_k的三阶语义图的顶点集,$\bar{W}_{M_k}$为归一化邻接矩阵.任意两个原始特征x_i与x_j关联的隐层特征表示z_i与z_j的余弦矩阵为$\hat{S}=\frac{z_i^T z_j}{\|z_i\|\|z_j\|}$,$\hat{S}\in R^{|n|\times|n|}$.通过重构多个不同模体$M_k$下的三阶语义图结构信息,短文本不同词节点之间的多种局部情感语义关系可以在三阶语义图结构得到体现[40].

为了更好地将不同模体的三阶语义图与短文本的情感语义融合,我们在预训练过程中对不同模体下三阶语义图中的边信息进行了调整:(1)若选择的模体M_k中只存在单向边,则删除三阶语义图结构中所有双向边;(2)若选择的模体M_k中只存在双向边,则删除三阶语义图结构中所有单向边;(3)若选择的模体中同时存在单向边和双向边,则三阶语义图结构中所有边的方向保持不变.单向边或者双向边通常不直接用于捕获文本之间的局部情感语义,但可以在一定程度上反映不同模体的上下文或者局部情感关联性.此外,在三阶语义图的预训练中,任何孤立的词节点(即没有与任何边相连的词节点)也会被删除,这主要是为了更好地反映三阶语义图的情感语义,同时降低三阶语义图增广的复杂性.

3.3 三阶语义图数据增广算法

现有的文本增广方法[41]容易引入新的无效噪声数据,可能使类别之间的边界变得更加不明显.因此,Luque[42]借鉴演化学习中的交叉思想,提出基于用例文本交叉增广的平衡策略以生成新的短文本数据.该方法可以在一定程度上使新的短文本数据的情感语义变化保持在合理的范围内,但对于提高短文本的情感识别能力仍然存在一定的限制.为避免文本增广过程中冗余噪声数据的产生,进一步提升少数类情感类别的识别能力,本文提出一种基于三阶语义图数据的增广算法,以解决短文本数据中存在的类不平衡问题.该方法从多个三阶模体的角度生成包含更多的局部情感语义特征的少数类样本.与传统的文本交叉增广[41]不同的是,本文在采用三阶语义图数据交叉增广方法构造新数据集的新文本数据这一过程中,只需要在相同情感标签类别的短文本数据集中生成三阶语义图结构的新数据,不需要额外的训练.三阶语义图数据交叉增广生成的短文本并不是通过简单的交换得到的,而是需要考虑不同模体下词节点依赖的情感语义差异性,从而避免生成新的噪声数据.

如图6所示,本文选择具有积极类情感标签的印尼短文本"T1: Taman anda sangat indah anda luar biasa"和"T2: Anjing itu sangat lucu dan itu kucing suka bermain bola",按照短文本中的先后顺序对词节点集进行标号,提取模体M_5下两个三阶语义图数据$G^{(5)}(s)$和$G^{(5)}(t)$的情感语义子图结构特征(图6(a)),然后采用最小割算法[43]分解(图6(b))和二部图匹配[44](图6(c)和图6(d))重组的方法得到不同模体下新三阶语义图结构的短文本数据$T1_{new}(M_5)$和$T2_{new}(M_5)$.为尽量保留原数据中基于模体M_5子图的完整情感语义信息,我们采用谱聚类[43]分别对两个三阶语义图文本数据$G^{(5)}(s)$和$G^{(5)}(t)$进行二分解.在分解过程中,将短文本的起始符和结束符定义为源点和汇点,基于模体M_5的邻接矩阵值作为三阶语义图数据中对应的边权值.为了确保在模体M_5下交叉增广以生成的新三阶语义图数据中能够蕴含更多完整的子图情感语义,我们采用基于二部图匹配的方法重组方法分别对分解后的$G^{(5)}(s1)$和$G^{(5)}(t2)$,$G^{(5)}(t1)$和$G^{(5)}(s2)$进行交叉增广生成新的三阶语义图结构数据$G^{(5)}(new_1)$和$G^{(5)}(new_2)$.以$G^{(5)}(s1)$和$G^{(5)}(t2)$交叉增广的过程为例,词节点(indah, 4)和(anda, 2)为一个子集,词节点(kucing, 6)和(suka, 7)为另一个子集.假设不考虑同一个子集内词节点的相关性,设词节点语言特征向量的余弦相似度为两个子集词节点的权值,则两个割中边集对应的顶点匹配问题就是一个二分

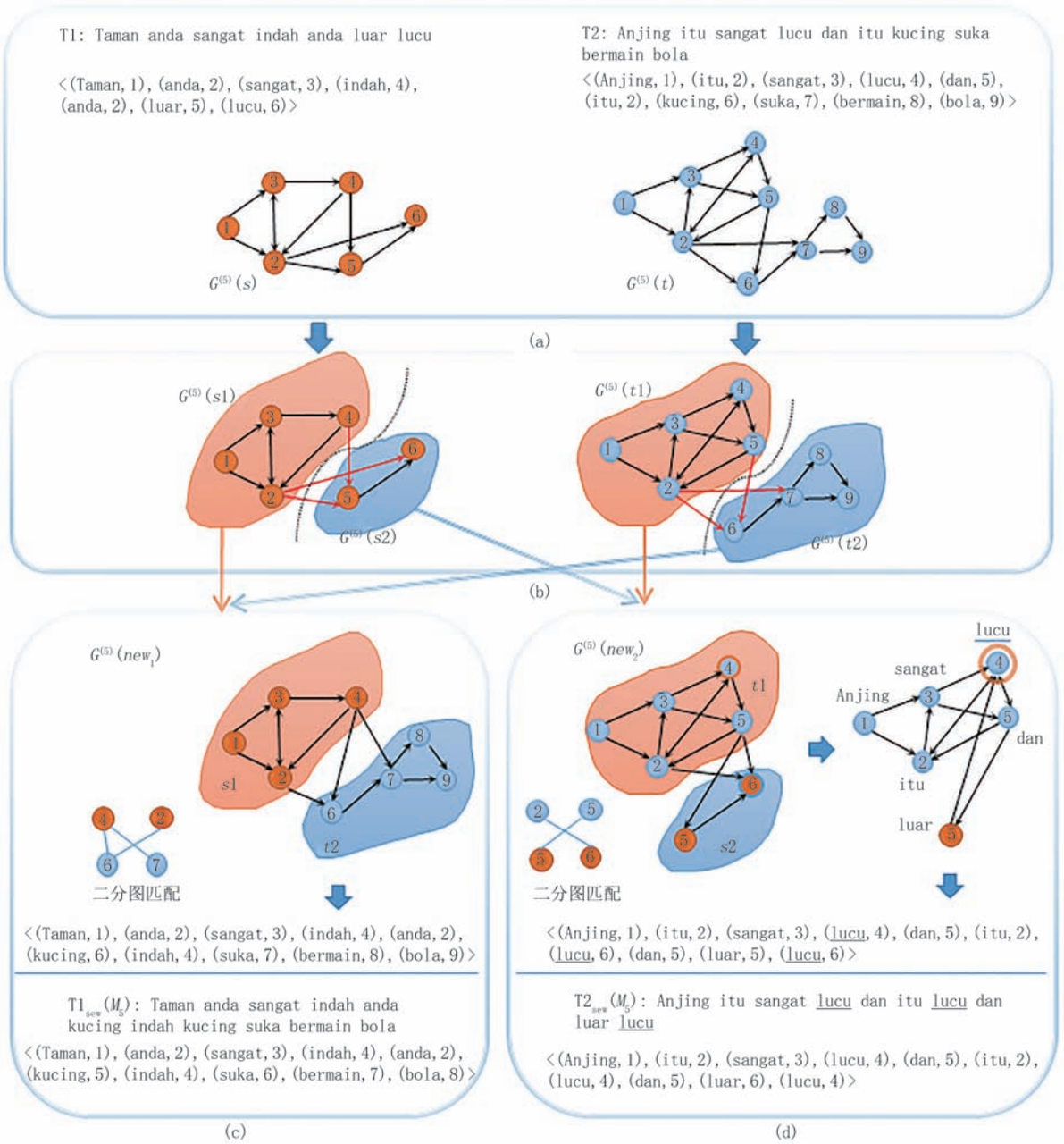


图6　基于三阶语义图数据增广算法生成新数据实例

图匹配问题. 我们采用KM算法[44]获得词节点集(indah，4)、(anda，2)和词节点集(kucing，6)、(suka，7)之间的最佳匹配关系. 若新增的边在至少1个M_5的模体中出现过，则添加相应的边以及边方向. 如图6(c)和图6(d)所示，通过分解和重组的方法交叉增广得到模体M_5下的新三阶语义图结构数据$G^{(5)}(new_1)$和$G^{(5)}(new_2)$，可生成对应的短文本数据“$T1_{new}(M_5)$: Taman anda sangat indah anda kucing indah kucing suka bermain bola”和“$T2_{new}(M_5)$: Anjing itu sangat lucu dan itu lucu dan luar lucu”，并根据增广文本对应的词节点序号向量短文本中的先后顺序分别对词节点集进行重新标号. 需要注意的是，在词节点集中的重组边的方向需要与当前选择模体M_5对应，如图6(c)中，在确定前序节点集(kucing，6)的基础上，若要使节点集(indah，4)、(kucing，6)、(suka，7)之间满足模体M_5的语义关系，只能是词节点集(indah，4)指向节点集(suka，7)之间的边关系，不能是节点集(suka，7)指向节点集(indah，4)的边关系. 此外，在图6(d)中增广生成的$G^{(5)}(new_2)$存在重叠的词节点“*lucu*”($G^{(5)}(t_1)$中的节点4与$G^{(5)}(s_2)$中的节点6)，因此，需要对重复的词节点进行合并操作. 通过在不同模

体下采用分解和重组的方法交叉增广，可以使生成的三阶语义图数据尽量包括完整的局部情感语义特征.

算法1. 三阶语义图数据增广算法

输入：原三阶语义图结构数据集 $Data=\{G_M(i)\mid 1\leqslant i\leqslant N\}$，情感类别标签集 $Y=\{y_i\mid 1\leqslant i\leqslant m\}$，不平衡率参数$\rho$

输出：新的三阶语义图结构数据集$NewData$

1. 按照情感类别标签集 Y 将不平衡数据集 $Data$ 分为 m 个子数据集，并分别统计每个数据集 $Data_j$ $(1\leqslant j\leqslant m)$ 的样本数 $n_j=num(Data_j)$;
2. 每个子数据集的不平衡率 $r_j=\dfrac{num(Data_j)}{\sum_{j=1}^{m}num(Data_j)}$;
3. for $1\leqslant$j$\leqslant m$
4. $newnum_j=\rho\times\sum_{j=1}^{m}num(Data_j)-num(Data_j)$;
5. $Tempnum=0$;
6. if $r_j<\rho$&& $Tempnum<\left\lceil\dfrac{newnum_j}{2}\right\rceil$
7. 随机生成两个不同的整数s和t, $s,t\in[1,N]$;
8. while $1\leqslant k\leqslant 13$&& $s\neq t$
9. 选取三阶模体M_k下两个三阶语义图结构信息的样本$G^{(k)}(s)$和$G^{(k)}(t)$;
10. if $G^{(k)}(s)\neq 0$ && $G^{(k)}(t)\neq 0$
11. 采用谱聚类将三阶语义图结构$G^{(k)}(s)$和$G^{(k)}(t)$分别进行二分解;
12. 对分解后的子图进行二分图匹配的重组，增广生成两个新的三阶语义图数据$G^{(k)}(new_1)$和$G^{(k)}(new_2)$;
13. 对$G^{(k)}(new_1)$和$G^{(k)}(new_2)$中重叠的词节点分别进行合并操作;
14. $G^{(k)}(new_1)=\varnothing$, $G^{(k)}(new_2)=\varnothing$;
15. end if
16. $G_M(new_1)=\{G^{(k)}(new_1)\mid 1\leqslant k\leqslant 13\}$;
17. $G_M(new_2)=\{G^{(k)}(new_2)\mid 1\leqslant k\leqslant 13\}$;
18. if $G_M(new_1)\neq\varnothing$ or $G_M(new_2)\neq\varnothing$
19. $Data_j=Data_j\cup G_M(new_1)\cup G_M(new_2)$;
20. $Tempnum=Tempnum+2$;
21. end if
22. end while
23. end if
24. $NewData=\bigcup_{j=1}^{m}Data_j$
25. end for
26. return $NewData$

三阶语义图数据增广方法的具体过程如算法1所示. 在算法1中，设初始的不平衡数据集为$Data$，三阶语义图数据的增广算法得到的新数据集为$NewData$. 如图5所示，三阶语义图交叉增广过采样方法主要分为两个阶段：

(1)按照情感类别将原不平衡数据集$Data$分为m个子数据集$Data_j$，并定义子数据集$Data_j$的不平衡率为 $r_j=\dfrac{num(Data_j)}{\sum_{j=1}^{m}num(Data_j)}$，其中$num(Data_j)$为子数据集$Data_j$的样本数，$max_num_data$为规模最大的子数据集样本数. ρ定义为多个情感类别的不平衡率参数，用于调节少数数据增加的短文本数量. 数据集$Data_j$的不平衡率越小，则新生成的三阶语义图结构短文本数据的数量越多.

(2)采用三阶语义图结构交叉增广少数类短文本数据. 在同一个少数类情感标签类别数据集$Data_j$中，随机选取三阶语义图短文本数据进行交叉合并得到新的$\rho\times\sum_{j=1}^{m}num(Data_j)-num(Data_j)$个三阶语义图数据. 若生成的三阶语义图数据中存在重复的词节点，则需要对其进行合并操作，重叠词节点在模体内的关联边关系保持不变，合并后的词节点特征为重叠词节点的平均值. 通过增广可以得到13个不同模体下的三阶语义图，每个三阶语义图都蕴含不同的某种局部情感语义信息，其对应的词节点个数以及边信息也不一致的，因此表示的局部情感语义有差异.

3.4 情感分类模型训练

为了融合不同模体下的短文本的三阶语义图情感信息，本文采用多图核学习的方法训练出三阶语义图结构信息的情感分类模型，如式(2)所示：

$$f(G_M)=sign\Big(\sum_{i=1}^{N}\alpha_i y_i\sum_{l=1}^{13}\beta_l K^{(l)}\left(G_i^{(l)},G^{(l)}\right)+b\Big) \tag{2}$$

其中$G_M=(G^{(1)},G^{(2)},\cdots,G^{(13)})$，表示测试集的三阶语义图结构样本. $K^{(l)}(G_i^{(l)},G^{(l)})$指模体$M_l$下不同三阶语义图结构的最小路径图核[45]，可以通过测试集中的新三阶语义图样本数据和训练集中每个三阶语义图样本数据计算得到. $\sum_{l=1}^{13}\beta_l K^{(l)}(G_i^{(l)},G^{(l)})$是不同模体下三阶语义图结构信息核函数的线性融合，需满足融合系数之和为1，即$\sum_{l=1}^{13}\beta_l=1$. α_i和b是待优化参数. y_i为三阶语义图数据所对应的短文本情感类别标签. 本文采用多图核学习的方法将不同模体下的三阶语义图结构信息映射到高维空间以进行融合，训练不平衡短文本的情感分类模型并得到分类结果.

3.5 计算复杂性

本文提出的TSGA模型的计算复杂度主要由

三个步骤组成：(1)在短文本的三阶语义图表示过程中，Seg-BERT[39]预训练过程的词表大小、词节点向量的维度以及隐层层数都是给定的参数，本文只考虑与短文本长度(三阶语义图表示的词节点n)有关的计算复杂度，即$O(c_1n+c_2)$. c_1与c_2为常量，与预训练过程中的模型参数设置有关. (2)三阶语义图信息增广算法的计算复杂度为$O(mk(l_V{l_E}^2))$. m为情感类别数，k为三阶模体的个数，l_V和l_E分别为短文本的三阶语义图表示的顶点节点和边数量. (3)多图核SVM分类模型训练的计算复杂度为$O(m(l_V+l_E))$，其中(l_V+l_E)为图核学习的计算复杂度[34]. 因此，本文提出的TSGA模型的总计算复杂度近似为$O(mk(l_V{l_E}^2)+ml_V+ml_E+c_1n)$.

4　实验与分析

本节将TSGA与其他传统的机器学习方法、深度网络模型以及文本增广等方法进行比较，从算法的分类性能、情感识别能力、参数敏感性分析等方面进行实验结果分析，验证本文提出的三阶语义图数据增广方法在不平衡文本情感多分类问题上的有效性.

4.1　实验集与评价指标

为了验证基于三阶语义图数据增广算法在不平衡情感分类问题上的可行性与有效性，本文在印尼语、马来语、英语以及中文四个公开的不平衡数据集上进行短文本多分类的实验验证. 为了验证TSGA模型在低资源语言短文本情感分类的性能，本文选取印尼语Ferdiana&Solihah[2-3]数据集、马来语Mohd Akid Haziq数据集①作为实验数据. 这些数据集的数据类别存在一定的不平衡性，某些短文本中还可能混有少量其他语言的词，因此传统的词向量表示无法对短文本特性进行精确表示. 印尼语Ferdiana&Solihah数据集是由两个规模较小的Solihah[2]数据集和Ferdiana数据集[3]合并得到的. 这两个公开的Twitter评论数据集都是跟印尼选取相关的评论，且规模都比较小. Mohd Akid Haziq数据集来自马来语商品和服务税相关的公开Twitter评论数据集. 此外，我们也选取英语SemEval2017数据集②以及中文的酒店评论数据集作为实验数据. 英语SemEval2017数据集是2017年语义评论比赛任务4中子任务A中的数据集，中文的酒店评论数据集③是由谭松波老师收集整理的数据集. 印尼语、马来语以及英语数据集的情感标注包含积极的、中性的和消极的三种情感级性，中文数据集中包含积极的和消极的两种情感级性. 不平衡短文本数据集的情感类别统计信息如表2所示.

表2　不平衡短文本数据集的统计信息

数据集	积极的	消极的	中性的	共计
Ferdiana&Solihah（印尼语）	2825	3205	5931	11 961
Mohd Akid Haziq（马来语）	1253	5051	2525	8829
SemEval2017（英语）	7059	3231	10 342	20 632
酒店评论数据集（中文）	7000	3000	---	10 000

不平衡短文本情感分类问题的混淆矩阵如表3所示，C_1、C_2和C_3分别表示情感类别标签为积极的、消极的和中性的. 基于混淆矩阵，不平衡短文本情感分类的总体准确率为

$$Accuracy=\frac{\sum_{i=1}^{3}\boldsymbol{n}_{ii}}{\sum_{j=1}^{3}\sum_{i=1}^{3}\boldsymbol{n}_{ij}} \tag{3}$$

表3　三分类混淆矩阵

		Predicted Class		
		积极的	消极的	中性的
True Class	积极的	$\boldsymbol{n}_{11}$	$\boldsymbol{n}_{12}$	$\boldsymbol{n}_{13}$
	消极的	$\boldsymbol{n}_{21}$	$\boldsymbol{n}_{22}$	$\boldsymbol{n}_{23}$
	中性的	$\boldsymbol{n}_{31}$	$\boldsymbol{n}_{32}$	$\boldsymbol{n}_{33}$

而准确率$Accuracy$并不能准确反映少数类短文本的情感分类性能. 因此，在三分类不平衡短文本分类问题中，通常还需要提高少数类情感的识别能力($F-measure$)或者平衡不同类情感的识别能力($G-means$). 在不平衡短文本情感分类中，对某一类的$F_i-measure$评价指标为

$$F_i-measure=\frac{2Recall_i\times Precision_i}{Recall_i+Precision_i} \tag{4}$$

$$Recall_i=\frac{\boldsymbol{n}_{ii}}{\sum_{j=1}^{3}\boldsymbol{n}_{ij}} \tag{5}$$

$$Precision_i=\frac{\boldsymbol{n}_{ii}}{\sum_{j=1}^{3}\boldsymbol{n}_{ji}} \tag{6}$$

① https://github.com/AkidFhmi/GSTMalaysia1218
② https://download.csdn.net/download/qq280929090/9818883
③ https://blog.csdn.net/noter16/article/details/75340354

式中$Recall_i$与$Precision_i$分别为C_i类情感的召回率和精度率. 为了平衡不同类别情感的类别识别能力并反映分类模型的整体性能,定义评价指标*G-means*为积极、消极以及中性三类情感召回率的几何平均为

$$G-means=\prod_{i=1}^{3}Recall_i^{\frac{1}{3}} \tag{7}$$

4.2 实验参数设置

在本文的实验中,由于印尼语、马来语短文本中含有少量的英语词节点信息,在Seg-Bert预训练过程中都采用混合语言的Bert模型[①]来获取词节点输入特征的嵌入表示. 考虑到词图结构的复杂性问题,本文考虑的窗口为3的短文本词图所对应的三阶语义图模型表示,因此模体的阶数也设置为3. 在多核SVM分类模型中执行十折交叉验证,返回不平衡短文本情感分类的准确率,TSGA模型使用的相关参数如表4.

表4 参数设置

超参数	值
词图窗口	3
模体的阶数	3
不平衡率参数	1
词节点嵌入表示的维度	128

4.3 对比实验

为了验证TSGA在不平衡短文本分类问题上的性能,本文在印尼语、马来语、英语以及中文的数据集上将TSGA模型与其他的情感分类模型进行实验对比. 以下是对各个模型的简单介绍.

(1)SVM[14]. 对短文本进行词向量表示,采用SVM完成情感模型的训练和分类. SVM是较传统的机器学习模型,在短文本情感分析任务中取得了较好的分类结果.

(2)DBN[46]. 情感词典与多个机器学习方法相结合的方法. 通过情感词典获取文本的语义情感特征,集成多个分类器的方法以提高短文本情感的分类准确性.

(3)SetConv[47]. 采用集合卷积操作对每个类别的短文本数据提取关键特征信息,来训练不平衡类别的多分类器,并提升情感分类性能.

(4)BiLSTM-CNN [21]. 首先用双向LSTM用于文本序列的表示任务,然后将卷积神经网络和双向LSTM相结合的传统深度学习模型用于短文本情感分类任务.

(5)C-BERT[48]. 在短文本进行BERT向量初始化的基础上,引入基于条件的掩码模型任务,提升基于上下文文本数据增广的短文本分类性能.

(6)Seg-BERT[39]. 对短文本进行统一的有效图实例表示,然后采用无监督Graph-BERT[38]预训练并微调的方式完成短文本情感分类任务.

(7)HGAT[24]. 通过局部图传播获取当前节点到不同相邻节点的重要情感语义信息,并采用图注意力网络表示的方法提高短文本情感分类性能.

(8)SP-GKL[34]. 对短文本进行最短路径词图表示,然后采用基于图核学习的支持向量机分类方法,实现短文本的情感分类.

(9)TU[7]. 将优化指标纳入数据的采样过程中,采用强化学习训练数据采样器,然后选择用例丢弃或者保留的欠采样方法,提升分类器的性能.

(10)LDAM[49]. 该方法用重采样技术来避免产生类别偏见,并强化少数类数据的公平感知策略,提升情感多分类性能.

(11)DARE[42]. 基于实例交叉增广的短文本分类方法. 该方法对短文本进行BERT初始化向量,借鉴演化算法交叉操作的思想,对同类别的文本进行交叉操作,增广生成新的短文本数据.

(12)TF-IGM-CW[8]. 基于文本增广过采样的不平衡情感分类方法. 在该方法中的平衡数据策略主要分为两个阶段:首先采用词替换的方法增加短文本数据的噪声样本,其次对噪声样本情感特征执行修正策略.

4.4 实验结果与分析

4.4.1 算法性能比较分析

为了验证TSGA在不平衡短文本多分类问题上的有效性,本文在印尼语、马来语、英语以及中文数据集上将提出的TSGA与其他10个分类模型进行对比,实验结果如表5~表8所示. 表5~表8描述了不同模型在印尼、马来语、英语以及中文数据集上的不平衡短文本情感分类对比实验结果,其中粗体表示在实验中最佳的评价指标值. 由表5~表8可知,本文提出的TSGA模型在4个不同语言的数据集上取得了比其他分类模型总体上更好的分类结果.

如表5所示,在印尼不平衡短文本数据集上,TSGA在情感类别识别方面表现出比其他模型更好的性能. 例如,在三类不同情感类别上TSGA模

① https://huggingface.co/xlm-roberta-base

型的*F-measure*值上都在72%以上. 与SVM、DBN以及SetConv模型相比，TSGA模型在准确率上均要高出8.25%以上. 与TF-IGM-CW、TU、LDAM以及DARE模型相比，TSGA模型可以在印尼语短文本少数类（积极的、消极的）上获得更高的F_i—*measure*值，这可能是由于三阶语义图文本增广生成的新数据可以包含更完整的情感语义特征，不存在跨标签短文本的噪声数据生成，从而提升了基于文本增广的平衡策略性能. 此外，与BiLSTM-CNN模型和C-BERT模型相比，TSGA具有更高的准确率，这主要是因为在印尼语短文本情感类别分布不平衡的情况下，TSGA获取的情感文本语义特征具有一定的局部多样性. 虽然SP-GKL、HGAT以及Graph-BERT模型可以获取除短文本语言特征外的词图结构情感语义信息，但其*G-means*值和准确率*Accuracy*还是比TSGA模型要低，这可能是因为情感类别数据之间的不平衡性导致训练的过程容易偏向多数类短文本. 同时，与TF-IGM-CW模型相比，TSGA模型可以在印尼语短文本少数类（积极的、消极的）上获得更高的*F-measure*值，这可能是由于三阶语义图文本增广生成的新数据可以表达不同模体的局部情感语义以及词节点之间的依赖关系，因此包含了更完整的语义以及情感特征，从而提升文本增广过采样的平衡策略性能.

表6～表8显示的是TSGA模型与其他分类模型在马来语、英文以及中文数据集上的不平衡短文本分类实验结果. 与其他模型对比，本文提出的

表5　不同模型在印尼语数据集上的不平衡短文本情感分类对比实验结果

模型	*F-measure*			*G-means*	*Accuracy*
	F_1-*measure*	F_2-*measure*	F_3-*measure*		
SVM	0.5032	0.5315	0.6091	0.6332	0.6872
DBN	0.5908	0.5309	0.6453	0.6628	0.7218
SetConv	0.6801	0.7023	0.7312	0.7120	0.7291
BiLSTM-CNN	0.6728	0.6881	0.6549	0.7235	0.7732
C-BERT	0.7031	0.6910	0.6632	0.7424	0.7817
Graph-BERT	0.7132	0.7024	0.6741	0.7523	0.7928
HGAT	0.6890	0.7137	0.6751	0.7321	0.7915
SP-GKL	0.6875	0.7032	0.6971	0.7512	0.8031
TU	0.6432	0.6835	0.6843	0.7502	0.7694
LDAM	0.6301	0.5902	0.6028	0.6291	0.7112
DARE	0.6731	0.6718	0.7017	0.7321	0.8017
TF-IGM-CW	0.7091	0.7336	**0.7519**	0.7532	0.7838
TSGA	**0.7307**	**0.7683**	0.7239	**0.7828**	**0.8116**

表6　不同模型在马来语数据集上的不平衡短文本情感分类对比实验结果

模型	*F-measure*			*G-means*	*Accuracy*
	F_1-*measure*	F_2-*measure*	F_3-*measure*		
SVM	0.5428	0.5810	0.5390	0.5331	0.6280
DBN	0.5559	0.6117	0.5690	0.6533	0.6834
SetConv	0.6313	0.6612	0.6171	0.7112	0.7621
BiLSTM-CNN	0.5630	0.6510	0.5876	0.6783	0.7634
C-BERT	0.5234	0.7120	0.5943	0.6342	0.7714
Graph-BERT	0.5874	0.7342	0.6321	0.6579	0.7857
HGAT	0.6702	**0.7665**	0.6432	0.6831	0.7518
SP-GKL	0.7315	0.6871	0.6720	0.7063	0.8072
TU	0.7210	0.6432	0.6345	0.6342	0.7817
LDAM	0.6491	0.6523	0.6198	0.6912	0.7021
DARE	0.7142	0.7321	0.7042	0.7218	0.7907
TF-IGM-CW	0.7215	0.7630	0.7315	0.7197	0.8167
TSGA	**0.7387**	0.7601	**0.7523**	**0.7321**	**0.8275**

表7　不同模型在英语数据集上的不平衡短文本情感分类对比实验结果

模型	*F-measure*			*G-means*	*Accuracy*
	F_1-*measure*	F_2-*measure*	F_3-*measure*		
SVM	0.6549	0.6356	0.7018	0.7306	0.7318
DBN	0.7013	0.7291	0.7321	0.7839	0.7935
SetConv	0.7701	0.8019	0.8212	0.8123	0.8997
BiLSTM-CNN	0.6856	0.8134	0.8366	0.7932	0.8192
C-BERT	0.6532	0.7342	0.8123	0.7921	0.8829
Graph-BERT	0.7234	0.7542	0.8235	0.8082	0.8271
HGAT	0.6892	0.8035	0.7932	0.7908	0.8236
SP-GKL	0.7265	0.8269	0.8212	0.8321	0.8563
TU	0.7143	0.7832	0.7942	0.8031	**0.9015**
LDAM	0.7421	0.8034	0.8198	0.8231	0.8721
DARE	0.7939	0.8207	0.8145	0.8291	0.8318
TF-IGM-CW	**0.7972**	0.8299	**0.8376**	0.8193	0.8920
TSGA	0.7847	**0.8321**	0.8239	**0.8330**	**0.9015**

TSGA模型在三个不平衡短文本数据集上取得了比较好的分类效果. 在马来语数据集上，TSGA模型的*Accuracy*和*G-means*值分别为82.75%和73.21%，比TF-IGM-CW模型、DARE模型、LDAM、TU模型、SP-GKL模型、HGAT模型、BiLSTM-CNN模型以及SetConv模型都分别提升了1.08%和1.03%以上. 在英文数据集SemEval2017和中文酒店评论数据集上，TSGA模型在准确率Accuracy指标上均取得了最大值. 在英文数据集SemEval2017上，TSGA模型比其他对比

表8 不同模型在中文数据集上的不平衡短文本情感分类对比实验结果

模型	*F-measure*		*Accuracy*
	F_1-*measure*	F_2-*measure*	
SVM	0.7934	0.7891	0.7801
DBN	0.8371	0.8178	0.8291
SetConv	0.9052	0.9126	0.9115
BiLSTM-CNN	0.9201	0.9100	0.9051
C-BERT	0.9034	0.8234	0.8732
Graph-BERT	0.9221	0.8453	0.8945
HGAT	0.9182	0.9032	0.8991
SP-GKL	0.9244	0.9121	0.9163
TU	09246	0.9203	0.9158
LDAM	**0.9245**	0.9221	0.9209
DARE	0.8942	0.9108	0.9103
TF-IGM-CW	0.9218	0.9134	0.9200
TSGA	0.9213	**0.9230**	**0.9288**

模型G-means指标值都高，这可能是因为融合三阶语义图结构信息的文本增广方法对提升平衡每类情感的识别能力具有一定的优势.

4.4.2 三分类情感识别性能分析

为了验证TSGA模型在不同情感类别任务上的情感识别能力，本文在印尼语、马来语以及英语数据集上对不同分类模型的*G-means*值进行实验比较. 根据*G-means*的定义，如果分类模型获取了较大的*G-means*值，则说明模型在三个情感类别上的召回率较高，也就是平衡每类情感的识别能力较强. 如图7所示，与其他模型相比，TSGA模型在印尼语、马来语以及英文数据集上的*G-means*值提升3.87%、1.43%以及0.11%. 这可能是因为TSGA模型在不同三阶模体下生成的新样本数据对提升不同类别的情感识别能力比较有效，且三阶语义图的图结构信息对印尼语、马来语文本的局部情感特征表示更具优势.

在印尼语数据集上，不同分类模型在不同情感类别上的F_i-*measure*值的统计结果如图8所示. 随着训练次数的增加，11个模型在多数类(中性的)情感类别上的F_3-*measure*值不断提升，最终都稳定在一个比较小的数值区间内. 然而，TSGA模型在少数类(积极类、消极类)情感类别上的F_i-*measure*值比其他模型更具有优势. 如图8所示，SVM模型、DBN模型以及TU模型对少数类(积极的)的情感识别效果并不理想. SP-GKL模型的F_1-*measure*比SVM模型、DBN模型、SetConv模型和BiLSTM-CNN模型都更高，说明提取短文本的词图结构情感

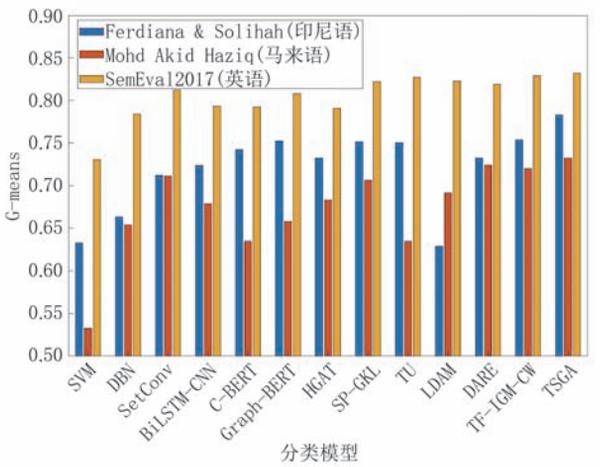
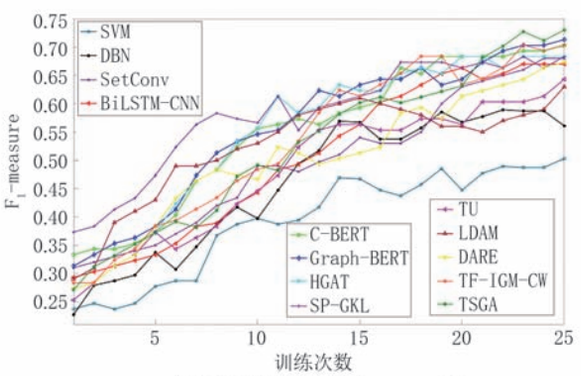
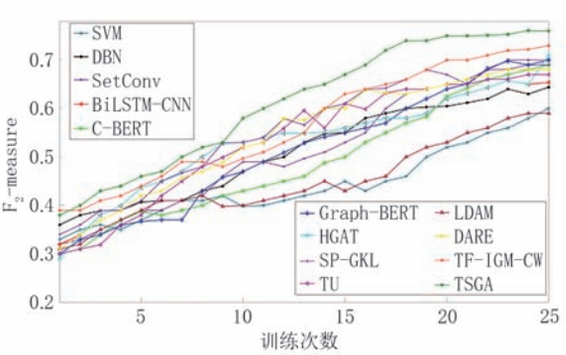


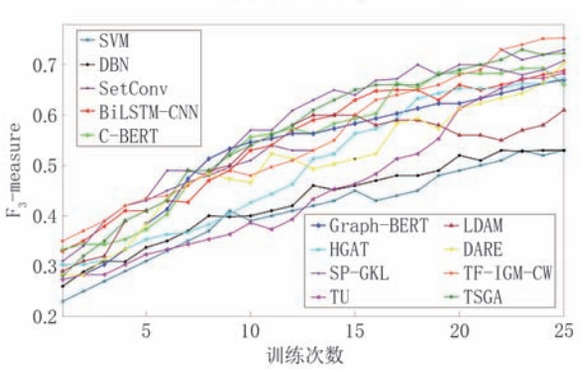
图7 不同分类模型的*G-means*值比较



(a) 积极类短文本的F_1-*measure*值



(b) 消极类短文本的F_2-*measure*值



(c) 中级类短文本的F_3-*measure*值

图8 不同分类模型在印尼语不平衡短文本 数据集上的情感类别

语义在一定程度上可以提升不平衡短文本分类任务的性能. 由图8可知，我们提出的TSGA模型在积极类情感上的F_1-*measure*值比SP-GKL模型、TF-IGM-CW模型和DARE模型分别高出约4%、2%和6%，这可能是因为融合三阶语义图结构信息的文本增广策略可以在没有精确的印尼语词向量的基础上生成包含完整情感语义特征的新数据，有效降低增广过程中引入新噪声的风险，还可以提升两个少数类（积极的、消极的）短文本的情感识别能力.

4.4.3　不平衡率参数ρ对TSGA的性能影响

不平衡率参数ρ是TSGA模型的一个重要参数，用于控制三阶语义图文本增广过程中生成少数类样本的数量. 当$\rho=\min r_j$时，TSGA模型只包含原始数据集中的短文本样本数据，为某类情感的不平衡率. 当$\min r_j<\rho<\max r_j$时，TSGA模型只对不平衡率最低的情感类别生成新样本. 当$\rho=\max r_j$时，所有情感类别的样本数达到平衡. 当$\max r_j<\rho<1$时，TSGA模型采用三阶语义图文本增广方法同时生成多个少数类样本. 当$\rho=1$时，所有情感类别的短文本样本数量不再增加. 为了进一步分析不平衡参数对TSGA模型的影响，我们将ρ的取值范围设置为[0.05, 1]，间隔为0.05，并记录TSGA模型的情感分类准确率变化过程.

在初始的短文本数据集中，若$\lfloor \min r_j*10 \rfloor/10 \geqslant 0.05$，则从样本数量最少的类中随机删除$\left\lfloor num(\boldsymbol{Data}_j)-0.05*\sum_{j=1}^{m} num(\boldsymbol{Data}_j) \right\rfloor$个少数类样本数据. 例如，为了确保初始的$\rho$取值为0.05，在印尼语数据集中，我们从少数类（积极类）中删除2227个数据.

TSGA模型在多个不平衡短文本数据集上的实验结果如图9所示. 随着少数类样本（消极类）数量的不断增加，TSGA模型在二分类中文数据集上可以获取较好的情感分类准确率. 具体而言，当$0.05<\rho<0.7$时，消极类样本不断增加，TSGA模型的准确率也不断增加；当$\rho=0.7$时，消极类和积极类的样本数据达到相同，TSGA模型的准确率几乎也达到最大值. 随着ρ的不断增加，消极类和积极类的样本都逐渐增加，但TSGA模型准确率基本保持不变. 此外，在英语数据集上，当$0.05<\rho<0.5$时，情感识别的准确率提升并不太明显，这可能是因为少数类（消极类）样本的占比数量太少的缘故. 当

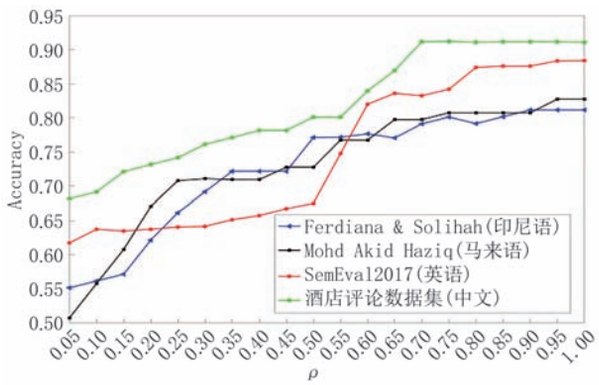


图9　参数ρ变化下TSGA模型的情感识别准确率

$0.50\leqslant\rho<0.80$时，随着积极类和消极类短文本样本的数量的不断增加，TSGA模型的总体准确率有所提升，这说明了增加少数类样本对提升多分类情感识别能力的有效性. 当$0.9\leqslant\rho<1$时，情感识别的准确率逐渐提高并趋于稳定. 在印尼数据集上，当$0.05<\rho<0.45$时，增加少数类（积极类）短文本样本提高了不平衡短文本情感分类准确率. 当$0.45\leqslant\rho<0.90$时，情感分类的准确率在小范围内有一些波动. 我们认为小范围的波动可能是由于印尼语数据集上的短文本中混有少量其他语言的词，导致生成的短文本具有不太清晰的情感语义特征. 当$0.90\leqslant\rho<1$时，情感识别的准确率基本保持不变. 总的来说，TSGA模型在印尼语、马来语、英语三个三分类数据集上表现出基本相似的准确率变化趋势. 随着有效少数类短文本数量的逐渐增加，不同情感类的边界可以得到有效拓展，从而使TSGA模型可以获得更准确的情感分类结果. 因此，有必要在短文本文本增广的过程中考虑不同模体的三阶语义图结构信息.

4.4.4　消融实验分析

为了验证短文本词图的三阶语义图表示与三阶语义图信息增广策略对不平衡短文本分类性能的影响，本文在印尼语、马来语、英语以及中文数据集上对TSGA分类模型上进行消融实验. 表9显示的是不同对比模型在初始不平衡率参数$\rho=0.05$的情况下的情感识别能力*G*¬*means*值和情感分类准确率*Accuracy*值上的实验结果. TSG和ISTGA分别指本文提出短文本词图的三阶语义图信息表示模块和三阶语义图数据增广策略模块. 模型GW_ISTGA和模型GAT_ISTGA分别指在采用传统的短文本词图表示方法和图注意神经网络（GAT）表示的基础上，结合本文提出的三阶语义图数据增广策略的

模型.在短文本词图的三阶语义图表示的基础上,模型TSG_EDA和模型TSG_DARE分别指采用传统的文本增广(EDA)和基于实例的短文本交叉增广(DARE)方法的平衡策略的分类模型.在TSGA_TSG模型中,我们对不平衡多分类数据只采用三阶语义图表示就直接进行分类训练,不采用高阶语义图数据增广方法.在模型TSG_ISTGA_SP-GKL和模型TSG_ISTGA_HGAT中,我们分别是将本文提出的三阶语义图表示方法以及增广方法,结合SP-GKL模型[34]和HGAT模型[24],并用于不同类别的情感分析任务.在实验中,我们将增广后的高阶语义图数据采用多个模体数据拼接的方法应用于SP-GKL和HGAT的分类模型.

从表9可以看出,在印尼语、马来语、英语以及中文数据集上,TSGA分类模型中的组件(TSG和ISTGA)在不平衡短文本的分类性能上存在一些差异.总体而言,TSG和ISTGA模块的引入明显提升了TSGA模型的分类性能.在印尼数据集上和马来语数据集上,采用词图表示和图注意力机制网络表示的方法替换本文提出的短文本词图的三阶语义图信息表示方法后,模型的分类性能和多分类情感识别能力都明显下降,这主要是因为TSGA可以通过不同的三阶模体获取更多不同子图的局部情感语义信息.此外,在英文数据集上,若采用传统的文本增广和基于实例的短文本交叉增广替换本文提出的三阶语义图信息增广平衡策略,不仅模型的情感分类性能有比较明显的下降,且多分类情感识别能力的*G-means*值也有所下降.TSGA分类模型在中文数据集上*Accuracy*指标上获得的值总体上优于其他模型,进一步说明TSGA分类模型中的TSG和ISTGA模块可以有效提升不平衡短文本的情感分类性能.此外,与TSGA_TSG模型的对比实验结果显示三阶语义图的数据增广方法在四个数据集上的*G-means*值和*Accuracy*指标均有不同程度的提升,进一步验证了本文提出的三阶语义图数据增广方法的有效性.

表10显示的基于高阶语义图增广方法生成数据应该在其他模型上的实验结果.通过对比,我们发现HGAT模型在四个数据集上的总体性能均劣于SP-GKL模型.通过在原HGAT模型基础上采用本文提出的高阶语义图生成数据,TSG_ISTGA_HGAT模型在印尼语、马来语以及英语三个数据集上获得的*G-means*值和*Accuracy*指标值都优于SP-GKL模型,且在印尼语和英语数据集上分别取得了81.01%和86.73%的准确率.与SP-GKL相比,TSG_ISTGA_SP-GKL模型的性能在4个数据集上均有明显提升,特别在印尼语、马来语以及英语数据集上均取得了较高的*G-means*值.这说明了基于高阶语义图增广生成的数据还可以应用到其他类型的图结构模型上提升情感分类的性能.

表9 初始不平衡率参数$\rho=0.05$的短文本情感分类消融实验结果

模型	Ferdiana &. Solihah (印尼语)		Mohd Akid Haziq (马来语)		SemEval2017 (英语)		酒店评论数据集 (中文)	
	G-means	*Accuracy*	*G-means*	*Accuracy*	*G-means*	*Accuracy*	*G-means*	*Accuracy*
GW_ISTGA	0.7012	0.7128	0.6217	0.6512	0.7912	0.7802	---	0.8256
GAT_ISTGA	0.7321	0.7432	0.6542	0.7234	0.7723	0.8219	---	0.8559
TSG_EDA	0.7454	0.7641	0.6943	0.7815	0.8219	0.8043	---	0.8856
TSG_DARE	0.7501	0.7932	0.7046	0.7934	0.8127	0.8421	---	0.8921
TSGA_TSG	0.7481	0.7929	0.6432	0.7865	0.8021	0.8363	---	0.8902
TSGA (Ours)	**0.7820**	**0.8109**	**0.7288**	**0.8204**	**0.8208**	**0.8803**	---	**0.9053**

表10 初始不平衡率参数$\rho=0.05$下基于高阶语义图增广方法生成数据的对比实验结果

模型	Ferdiana&.Solihah (印尼语)		Mohd Akid Haziq (马来语)		SemEval2017 (英语)		酒店评论数据集 (中文)	
	G-means	*Accuracy*	*G-means*	*Accuracy*	*G-means*	*Accuracy*	*G-means*	*Accuracy*
HGAT	0.7321	0.7915	0.6432	0.6831	0.7908	0.8236	---	0.8991
TSG_ISTGA_HGAT	0.7781	**0.8101**	0.6831	0.7321	0.8012	**0.8673**		0.9011
SP-GKL	0.7512	0.8031	0.6720	0.7063	0.8321	0.8563	---	0.9163
TSG_ISTGA_SP-GKL	**0.7812**	0.8091	**0.7147**	**0.7617**	**0.8327**	0.8646	---	**0.9147**

5　总结与展望

本文在短文本的三阶语义图模型表示的基础上，提出了基于三阶语义图数据增广算法，实现了不平衡短文本数据的情感类别平衡分布以及情感分类．该模型不仅可以将多个模体下的三阶语义图结构信息进行有效融合，还可以在没有精确词向量表示的基础上生成具有局部情感语义特征的少数类短文本数据，避免由于新增数据导致的类别重叠问题，从而提升少数类短文本的情感识别性能．实验结果表明，本文提出的TSGA分类模型在一定程度上提升了不平衡短文本的分类准确率，且在印尼语、马来语不平衡短文本数据集上较好地提升了对少数类短文本的情感识别能力．这主要是因为传统的文本增广方法在没有精确的词向量表示的情况下所生成的新数据容易产生过拟合的情况，而TSGA模型对语言特征的依赖性没那么强．本文仅考虑短文本数据的不平衡情感分类工作，这主要是考虑到在三阶语义图结构的信息表示下，过长的文本数据可能会对图的复杂度提出挑战．

在下一步工作中，我们将继续研究文本情感分类工作，尝试将三阶语义图数据增广策略用于解决长文本数据的不平衡学习问题．此外，语言资源不足问题、多文化下短文本的跨语言问题等都是不平衡短文本情感分析中需要考虑的问题[50]，因此低资源语言短文本的情感分析问题也是我们在未来工作中需要研究的问题．

参　考　文　献

[1] Hatef A, Luqiu L R. Where does Afghanistan fit in China's grand project? A content analysis of Afghan and Chinese news coverage of the "One Belt, One Road" initiative. International Communication Gazette, 2018, 80(6): 551-569

[2] Solihah R. Peluang dan tantangan pemilu serentak 2019 dalam perspektif politik. Jurnal Ilmiah Ilmu Pemerintahan, 2018, 3(1): 73-88

[3] Ridi Ferdiana W F, Purwanti D D, Ayu A S T, et al. Twitter sentiment analysis in under-resourced languages using byte-Level recurrent neural model. International Journal of Advanced Computer Science and Applications, 2019, 10(8): 1-5

[4] Goyal A, Rathore L, Kumar S. A Survey on solution of imbalanced data classification problem using SMOTE and extreme learning machine. Communication and Intelligent Systems, 2022, 7(5): 23-31

[5] Kubler S, Liu C, Sayyed Z A. To use or not to use: Feature selection for sentiment analysis of highly imbalanced data. Natural Language Engineering, 2018, 24(1): 3-37

[6] Cui Y, Jia M, Lin T-Y, et al. Class-balanced loss based on effective number of samples//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. California, USA 2019: 9268-9277

[7] Peng M, Zhang Q, Xing X, et al. Trainable undersampling for class-imbalance learning//Proceedings of the AAAI Conference on Artificial Intelligence. Hawaii, USA, 2019: 4707-4714

[8] Pang Z, Li H, Wang C, et al. A two-stage balancing strategy based on data augmentation for imbalanced text sentiment classification. Journal of Intelligent and Fuzzy Systems, 2021, 40(5): 10073-10086

[9] Song J, Huang X, Qin S and Song Q. A bi-directional sampling based on K-means method for imbalance text classification// Proceedings of the IEEE/ACIS 15th International Conference on Computer and Information Science. Okayama, Japan, 2016:1-5

[10] Cambria E, Das D, Bandyopadhyay S, et al. Affective computing and sentiment analysis. A Practical Guide to Sentiment Analysis. Berlin:Springer, 2017, 1-10

[11] Zeng X Q, Hua X, Liu P S, et al. Emotion wheel and lexicon based text emotion distribution label enhancement method. Chinese Journal of Computers, 2021, 44(6): 1080-1094. (in Chinese)
(曾雪强，华鑫，刘平生等．基于情感轮和情感词典的文本情感分布标记增强方法．计算机学报，2021，44(6)：1080-1094)

[12] Chekima K, Alfred R. Sentiment analysis of Malay social media text//Proceedings of the International Conference on Computational Science and Technology. Zürich, Switzerland, 2017: 205-219

[13] Lailiyah M, Sumpeno S, Purnama I E. Sentiment analysis of public complaints using lexical resources between Indonesian sentiment lexicon and sentiwordnet// Proceedings of the International Seminar on Intelligent Technology and Its Applications. Surabaya, Indonesia, 2017: 307-312

[14] Chen Z, Qian T Y, Li W L, et al. Low-resource aspect-based sentiment analysis: A survey. Chinese Journal of Computers, 2023, 16(07) (in Chinese)
(陈壮，钱铁云，李万理等．低资源方面级情感分析研究综述．计算机学报，2023)，16(07)

[15] Phienthrakul T, Kijsirikul B, Takamura H, et al. Sentiment classification with support vector machines and multiple kernel functions//Proceedings of the International Conference on Neural Information Processing. Vancouver, Canada, 2009: 583-592

[16] Fiarni C, Maharani H, Irawan E. Implementing rule-based and naive Bayes algorithm on incremental sentiment analysis system for indonesian online transportation services review// Proceedings of the 10th International Conference on Information Technology and Electrical Engineering. Yogyakarta, Indonesia, 2018: 597-602

[17] Sadanandan A A, Osman N A, Saifuddin H, et al. Improving accuracy in sentiment analysis for Malay language//Proceedings of the 4th International Conference on Artificial Intelligence and

Computer Science. Langkawi, Malaysia, 2016: 28-29

[18] Boiy E, Moens M-F. A machine learning approach to sentiment analysis in multilingual Web texts. Information Retrieval, 2009, 12(5): 526-558

[19] Cai, R, Hao, Z, Wen W, and Huang H. Kernel based gene expression pattern discovery and its application on cancer classification. Neurocomputing, 2010, 73(13-15): 2562-2570

[20] He Y X, Sun S T, Niu F F, et al. A deep learning model enhanced with emotion semantics for microblog sentiment analysis. Chinese Journal of Computers, 2017, 40(4): 773-790 (in Chinese)
(何炎祥, 孙松涛, 牛菲菲等. 用于微博情感分析的一种情感语义增强的深度学习模型. 计算机学报, 2017, 40(4): 773-790)

[21] Wang J H, Liu T W, Luo X, et al. An LSTM approach to short text sentiment classification with word embeddings// Proceedings of the 30th Conference on Computational Linguistics and Speech Processing. Hsinchu, China, 2018: 214-223

[22] Chen Y L, Fu Q K, Zhang Y. Applications of graph neural network for natural language processing. Journal of Chinese Information Processing, 2021, 35(03): 1-23 (in Chinese)
(陈雨龙, 付乾坤, 张岳. 图神经网络在自然语言处理中的应用. 中文信息学报, 2021, 35(03): 1-23)

[23] Yao L, Mao C, Luo Y. Graph convolutional networks for text classification//Proceedings of the AAAI Conference on Artificial Intelligence. Hawaii, USA, 2019: 7370-7377

[24] Hu L M, Yang T C, Shi C, Ji H Y, and Li X L. Heterogeneous graph attention networks for semi-supervised short text classification// Proceedings of the Conference on Empirical Methods in Natural Language. Hong Kong, China, 2019: 4821-4830

[25] Hajibabaee P, Malekzadeh M, Heidari M, et al. An empirical study of the graphsage and word2vec algorithms for graph multiclass classification//Proceedings of the 12th Annual Information Technology, Electronics and Mobile Communication Conference.Vancouver, Canada, 2021: 0515-0522

[26] Li X F, Li J, Dong Y F, et al. A new learning algorithm for imbalanced data-PCBoost. Chinese Journal of Computers, 2012, 35(2): 202-209 (in Chinese)
(李雄飞, 李军, 董元方, 等. 一种新的不平衡数据学习算法 PCBoost. 计算机学报, 2012, 35(2): 202-209)

[27] Feng S Y, Gangal V, Wei J, et al. A survey of data augmentation approaches for NLP. arXiv preprint arXiv: 2105.03075, 2021

[28] Liu S, Lee K, Lee I. Document-level multitopic sentiment classification of email data with bilstm and data augmentation. Knowledge-based Systems, 2020, 197: 105918

[29] Cheng Y, Jiang L, Macherey W, et al. Advaug: Robust adversarial augmentation for neural machine translation. arXiv preprint arXiv:2006.11834, 2020

[30] Yan G, Li Y, Zhang S, et al. Data augmentation for deep learning of judgment documents//Proceedings of the International Conference on Intelligent Science and Big Data Engineering. Nanjing, China, 2019: 232-242

[31] Chen Y, Wang J, Li P, et al. Single document keyword extraction via quantifying higher order structural features of word co-occurrence graph. Computer Speech & Language, 2019, 57: 98-107

[32] Zhao T, Liu , Günnemann S, et al. Graph data augmentation for graph machine learning: A survey. arXiv preprint arXiv: 2202.08871

[33] Hedderich M A, Lange L, Adel H, et al. A survey on recent approaches for natural language processing in low-resource scenarios. arXiv preprint arXiv:2010.12309, 2020.

[34] Nikolentzos G, Meladianos P, Rousseau F, et al. Shortest-path graph kernels for document similarity//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark, 2017: 1890-1900.

[35] Rossi R A, Ahmed N K, Koh E. Higher order network representation learning. Companion of the Web Conference. Lyon, France, 2018: 3-4

[36] Yan X, Huang H., Hao Z, and Wang J. A graph-based fuzzy evolutionary algorithm for solving two-echelon vehicle routing problems. IEEE Transactions on Evolutionary Computation, 2019, 24(1): 129-141

[37] Benson A R, Gleich D F, Leskovec J. Higher order organization of complex networks. Science, 2016, 353(6295): 163-166

[38] Chen X, Cai R, et al. Motif graph neural network. IEEE Transactions on Neural Networks and Learning Systems, 2023, 15(7):1-15

[39] Zhang J. Segmented GRAPH-BERT for graph instance modeling. arXiv preprint arXiv:2002.03283, 2020

[40] Zhang J, Zhang H, Xia C, et al. Graph-bert: Only attention is needed for learning graph representations. arXiv preprint arXiv: 2001.05140, 2020

[41] Bayer M, Kaufhold M-A, Reuter C. A survey on data augmentation for text classification. ACM Computing Surveys, 2022, 55(7):1-39

[42] Luque F M. Atalaya at tass 2019: Data augmentation and robust embeddings for sentiment analysis. arXiv preprint arXiv: 1909.11241, 2019

[43] Boykov Y, Kolmogorov V. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2004, 26(9): 1124-1137

[44] Zhu H, Zhou M, Alkins R. Group role assignment via a Kuhn-Munkres algorithm-based solution. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 2011, 42(3): 739-750

[45] Shervashidze N, Schweitzer P, Van Leeuwen E J, et al. Weisfeiler-Lehman graph kernels. Journal of Machine Learning Research, 2011, 12(77):2539-2561

[46] Al-Saffar A, Awang S, Tao H, et al. Malay sentiment analysis based on combined classification approaches and Senti-lexicon algorithm. PloS One, 2018, 13(4):e0194852

[47] Gao Y, Li Y F, alet, Setconv: A new approach for learning from imbalanced data. arXiv preprint arXiv:2104.06313,2021

[48] Wu X, Lv S, Zang L, et al. Conditional bert contextual augmentation //Proceedings of the International Conference on Computational Science. Faro, Portugal, 2019: 84-95

[49] Subramanian S, Rahimi A, Baldwin T, alet, Fairness-aware class imbalanced learning. arXiv preprint arXiv: 2109.10444, 2021

[50] Yan X, Huang H, Jin Y, et al. Neural architecture search via multi-hashing embedding and graph tensor networks for multilingual text classification. IEEE Transactions on Emerging Topics in Computational Intelligence. 2023, 8(1): 350-363

YAN Xue-Ming, Ph. D., associate professor. Her research interests include natural language processing, machine learning, and graph optimization.

HUANG Han, Ph. D., professor, Ph. D. supervisor. His research interests include microcomputing theory and methods, intelligent software engineering, and data intelligence engineering.

JIN Yao-Chu, Ph. D., professor, Ph. D. supervisor. His main research interests include data-driven evolutionary optimization of complex systems, evolutionary multi-objective machine learning, federated learning and secure machine learning, and evolutionary developmental systems and morphological developmental robotics.

ZHONG Guo, Ph. D., associate professor. His research interests include data mining and machine learning.

HAO Zhi-Feng, Ph. D., professor, Ph. D. supervisor. His main research interests include machine learning, computational intelligence, and algebra and combinatorial optimization.

Background

Most sentiment short text data generated on social media exhibit a certain degree of class imbalance. This imbalance often results from the disproportionate representation of certain sentiments—typically, positive or neutral emotions dominate, while negative or less common sentiments are underrepresented. Traditional methods for sentiment classification tend to prioritize the majority class because models trained on imbalanced data naturally focus on the most prevalent patterns. Consequently, these methods overlook the minority class, leading to poor performance in identifying less frequent but potentially significant emotions. However, it is crucial to focus on the minority class in sentiment classification, as correctly identifying emotional information within this class can be particularly valuable for tasks like detecting customer dissatisfaction, monitoring public opinion, and identifying emerging issues. In recent years, various strategies have been proposed to balance the distribution of data and enhance the performance of sentiment classification for imbalanced short texts. One such strategy is the use of text augmentation techniques, which generate additional short text data specifically for the minority class during the augmentation process. By artificially increasing the number of minority class examples, these techniques aim to provide a more balanced dataset for training classifiers. However, if the generated data fails to capture the complete semantic features of sentiment, it may lead to overlapping emotions across different categories. This overlap can confuse the classification model, causing it to misclassify sentiments and reducing overall accuracy.

In this study, we propose a method that leverages a third-order semantic graph for text augmentation to address imbalanced multi-class sentiment classification. This approach effectively generates minority short text data by designing a third-order semantic graph structure that expresses local emotional semantics. The third-order semantic graph captures complex relationships among words by considering not just direct word pairs but also the interactions involving intermediary words. This allows for a more nuanced representation of sentiment and word dependencies, ensuring that the augmented data accurately reflects the emotional context of the minority class. By utilizing this sophisticated graph structure, our method ensures a balanced distribution of sentiment classification across multiple classes for imbalanced short text data. To evaluate the effectiveness of our proposed approach, experiments were conducted on four public imbalanced short text datasets in Indonesian, Malay, English, and Chinese. Comparative experimental results demonstrate that our approach outperforms state-of-the-art algorithms in improving the recognition accuracy of minority emotions, particularly when the short text cannot generate an accurate word vector model. This work is supported by the National Natural Science Foundation of China (62276103, 62136003, 62476163), International Collaboration Fund for Creative Research Teams (ICFCRT) of NSFC (No. W2441019), Guangdong Basic and Applied Basic Research Foundation (2023B1515120020), and Innovation Team Project of General Colleges and Universities in Guangdong Province (2023KCXTD002).