# 融合汉越关联关系的多语言事件观点对象识别方法

**李格格**[1,2]，**郭军军**[1,2]，**余正涛***[1,2]，**相艳**[1,2]
1.昆明理工大学，信息工程与自动化学院，昆明，650500
2.昆明理工大学，云南省人工智能重点实验室，昆明，650500
1303717217@qq.com, guojjgb@163.com
ztyu@hotmail.com, sharonxiang@126.com

## 摘要

越南语观点对象识别是越南语事件观点分析的重要研究内容。由于汉越两种语言的语法结构上存在差异，使得多语言事件关联复杂，观点对象表征困难。现有研究方法仅能实现汉越双语的表征，未能有效捕获并利用汉越双语事件中要素的关联关系。因此，本文提出一种融合汉越关联关系的多语言事件观点对象识别方法，利用中文和越南语事件间的要素共现和整体语义关联构建汉越多语言事件表征网络，使用多语言预训练语言模型获得要素节点的特征向量，利用图卷积网络对节点信息进行聚合，得到同一语义空间下汉越双语的公共表征，实现汉越事件观点对象的识别。实验结果表明本文模型能够更有效地构建多语言关联信息，其F1值较多个基线模型都有明显提高。

**关键词：** 观点对象识别；多语言事件关联；图卷积网络

# A Multilingual Event Opinion Target Recognition Method Incorporating Chinese and Vietnamese Association Relations

**Gege Li**[1,2]，**Junjun Guo**[1,2]，**Zhengtao Yu***[1,2]，**Yan Xiang**[1,2]
1. Faculty of Information Engineering and Automation, Kunming University of Science and Technology Kunming 650500, China
2. Yunnan Key Laboratory of Artificial Intelligence, Kunming University of Science and Technology Kunming 650500, China
1303717217@qq.com, guojjgb@163.com
ztyu@hotmail.com, sharonxiang@126.com

## Abstract

Vietnamese opinion target recognition is an important research component of Vietnamese event opinion analysis. Due to the differences between the grammatical structure of Chinese and Vietnamese, it makes the association of multilingual events complex and the representation of opinion target difficult. Existing research methods can only realize the bilingual representation of Chinese and Vietnamese, and fail to effectively capture and utilize the association relationship of elements in Chinese and Vietnamese events. Therefore, this paper proposes a multilingual event opinion target recognition method that integrates Chinese and Vietnamese association relations, using element co-occurrence and overall semantic association between Chinese and Vietnamese events to build a network of Chinese and Vietnamese multilingual event representation, using a multilingual pre-trained language model to obtain the feature vectors of element nodes, and using graph convolutional network to aggregate node information to obtain a common representation of Chinese and Vietnamese in the same semantic space. In

*余正涛（通讯作者）：ztyu@hotmail.com
**基金项目**：国家自然科学基金（U21B2027，61972186，62266027，62266028）；云南省科技重大专项（202302AD080003，202103AA080015）；云南省基础研究计划项目（202301AS070047，202301AT070444）

order to achieve the recognition of Chinese and Vietnamese opinion target recognition. The experimental results show that the model in this paper can construct multilingual association information more effectively, and its F1 values are significantly improved compared with several baseline models.

**Keywords:** Opinion target recognition , Multilingual event correlation , Graph convolutional network

## 1 引言

互联网的快速发展推动了中越两国交流，从社交媒体评论文本中挖掘两国用户的观点，掌握用户对事件的关注，对处理好与越南的国际关系、区域经济发展和文化交流有着重要的作用，同时为政府及企业正确把握汉越舆情动态并及时做出应对措施提供有效保障。越南语标注数据资源的稀缺，阻碍了其观点对象识别方法的研究，可通过多语言观点对象识别（Multilingual Opinion Target Recognition）的方法，利用具有丰富标记数据的中文通过知识迁移帮助标注资源稀缺的越南语实现多语言观点对象识别任务。


Figure 1: “新冠疫情”数据集上的汉越社交媒体评论样例

如图1所示的是汉越“新冠疫情”数据集中有关“新冠疫苗”和“防疫措施”两种不同观点对象的评论句。图中左半部分描述的是汉越两国用户针对“新冠疫苗”观点对象发出的评论，右侧评论则是针对“防疫措施”的讨论，通过观察上图可以发现中文和越南语评论在针对同一事件时讨论的内容较为接近，关注的重点也较为相似，利用这种关联特征可以较好地捕获汉越双语评论的全局特征（汉越评论之间的关联关系）和局部特征（评论中关键词所携带的语义信息）。通过对关联关系和语义信息进行建模，能够得到信息互补的特征表示学习模型，从而较好地完成迁移任务，解决越南语标注资源稀缺的问题。

目前，在多语言观点对象识别的研究中，主要通过基于传统机器学习的方法和基于深度学习的方法进行观点对象识别，根据每个领域的评论表征来学习特定的观点对象分类器。基于传统机器学习的方法通过制定相关规则并融入领域相关信息等外部知识利用算法提升识别性能，基于深度学习的方法通过使用神经网络提取数据特征进行观点对象的识别。这些模型的训练过程需要大规模且高质量的标注数据集，但是在面对不同的应用场景时，构建这样规模的训练数据集成本较高，同时利用传统的特征编码模式只能考虑到单语语料库中各评论文本的局部特征，不能很好的做到多语言间的知识迁移。

针对以上问题，本文提出一种融合汉越关联关系的多语言事件观点对象识别方法，该方法通过将汉越社交媒体评论文本和其中的关键词(高频词)作为节点构建异构图，结合评论文本节点的输入表征，通过图卷积网络准确地捕获汉越双语评论间观点对象的关联信息，提高观点对象表征学习和模型识别性能。本文的主要贡献如下:

（1）在中文和越南语评论文本上利用异构图进行关联关系构建，通过构建多种类型的节点和边关系，捕捉各节点之间丰富的关系结构，得到汉越评论文本数据在同一嵌入空间下的对应关系。

（2）使用多语言预训练语言模型获取评论文本的特征向量，并将其作为评论文本节点的输入表征，使用图卷积网络学习节点特征并基于图结构迭代更新评论文本表征，进行汉越观点对象的识别。

（3）在所构建的汉越评论数据集上进行了实验，相比已有的基准模型，所提模型的性能都有较大的提升。

## 2 相关工作

观点对象是由带有情感偏见的情感词所修饰的对象，通常是社交网络中的一个特定主题，或者是电子商务平台上的一个特定产品或产品评论的一部分。观点对象识别是从预定义的标签集合中为评论文本分配对应的标签，观点对象识别策略可以分为以下两大类方法。

### 2.1 基于传统机器学习的方法

基于传统机器学习的方法主要是通过人工制定规则来分析语料或者通过融入领域相关信息等外部知识利用机器学习的算法提升识别性能，主要分为基于规则和统计的学习方法以及基于机器学习的方法。

#### 2.1.1 基于规则和统计的学习方法

基于规则和统计的学习方法主要对语料库进行分析，结合分析制定词性规则、词序列规则和句法规则。倪茂树and 林鸿飞 (2007)提出了一种利用关联规则和极性分析方法挖掘观点特征的算法，从而更好的识别出商品评论中观点对象的类别。Qiu et al. (2011)用情感词识别观点对象的修饰关系和整体的从属关系词，取得了良好的实验结果。这些方法非常依赖规则和具体语言，难以覆盖所有情况导致只适合小规模的数据，而且系统移植性不强，根据任务的不同需要设定新的规则，建设周期长并且代价比较高昂。

#### 2.1.2 基于机器学习的方法

基于机器学习的方法通过融入领域相关信息等外部知识利用机器学习的算法提升识别性能。Titov and McDonald (2008)采用多粒度的主题模型分析并识别文本中的观点对象，并在分析结果的基础上，归类出相同的观点对象，对相似度较大的观点对象进行聚类。Moghaddam and Ester (2011)利用狄利克雷分布（LDA）提取观点对象和相应的评级产品在线评论。Li et al. (2012)提出了一种新的关系自适应引导(RAP) 算法，通过利用标记的源域数据来获得主题词和观点对象之间的关系。Li et al. (2018)将每一个时间戳对应的观点对象特征与原始抽取的观点对象特征进行融合，另外利用观点对象识别过程中的坐标结构和抽取到的观点对象特征进行交互，通过探索到的两种信息提升观点对象识别模型的性能。基于机器学习的方法比起基于规则和统计的学习方法有一定的改进，但还是需要人工对文本特征进行标记，人为的主观因素会影响模型的性能，同时机器学习需要依赖先验知识的质量和大量的标记数据，执行的速度会比较慢，难以适应如今信息量爆炸的时代。

### 2.2 基于深度学习的方法

基于深度学习的方法通过训练神经网络，使用CNN、RNN和LSTM等各种典型的神经网络对观点对象的类别进行识别，Ding et al. (2017)提出使用规则在循环神经网络上生成辅助监督，以学习每个单词领域不变的隐式特征表示。Nguyen and Le Nguyen (2018)在SenTube数据集上提出了卷积N-gram BiLSTM词嵌入，用于进行多语言观点对象的识别。

这些模型会优先考虑文本的局部信息和顺序信息，能够很好的捕获连续词序列中的语义和句法信息，但是它们忽略了多语言的全局词共现，而全局词共现中携带了不连续以及长距离的

语义信息。随着深度学习的不断发展，近些年来图卷积网和多语言预训练语言模型被广泛应用到观点对象识别的任务当中。

### 2.2.1 基于图卷积网络的方法

图卷积神经网络可以通过在节点之间传递信息建立图模型，Yao et al. (2019)提出文本图卷积神经网络TextGCN，基于词共现和文档词关系建立单语言图，该方法可以同时学习单词和文档的嵌入。在多语言的任务中，Li et al. (2020)将图神经网络应用在元学习方面。Wang et al. (2021)提出CLHG模型，使用了类似于TextGCN方法的构图方式，将图神经网络用在多语言文本分类中，解决了原有模型只注重语义信息而忽略句法信息这一缺点。图卷积网络虽然擅长将图中的全局信息卷积成一个文本，但不能同时捕捉上下文的相关性与关联信息，剥离了文本中词与词之间的关联性，难以获得高效的性能。

### 2.2.2 基于多语言预训练语言模型的方法

在单语预训练语言模型的不断发展下，部分学者将焦点聚集在多语言预训练语言模型的训练中，并且训练出来的模型在许多下游任务中表现出优异的跨语言迁移能力。Kenton and Toutanova (2019)通过在104种语言的维基百科语料库上进行训练推出了mBert，在多语言迁移方面取得了不错的效果。Lample and Conneau (2019)提出的XLM模型，通过构造编码器编码多种语言的句子到同一嵌入空间来增强多语言的共享词汇。Conneau et al. (2020)在之前的基础上推出了多语言预训练语言模型XLM-R，增加了模型的语言数量和训练示例的数量。尽管多语言预训练语言模型已经取得了许多最新的成果，这些模型没有明确考虑语言之间的句法差异，导致目标语言的泛化性能下降，同时任务特定的结构依赖问题为模型性能的进一步提高带来了许多限制。

## 3 模型介绍

本文提出一种融合汉越关联关系的多语言事件观点对象识别方法，不仅关注汉越双语评论之间语义差异的问题，同时也关注汉越双语评论中观点对象之间的对齐关系，模型总体架构如图2所示，它主要包含四个网络：节点特征表示、多语言异构图的构建、节点特征学习和观点对象类别预测。

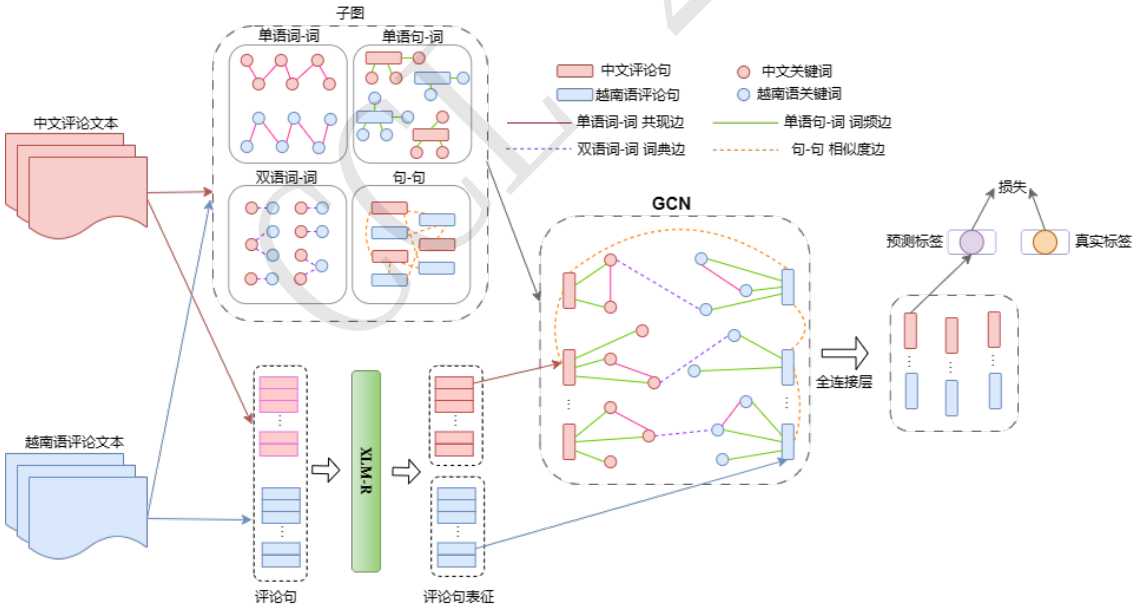


Figure 2: 融合汉越关联关系的多语言事件观点对象识别模型

使用多语言预训练语言模型获取汉越社交媒体评论文本的节点特征，将评论文本和其中的关键词作为异构图的节点，并基于评论文本中词共现、词对齐、词频信息和语义相似度的关系构边，利用图卷积网络对节点特征进行学习，并对节点进行线性转换输出评论文本节点的预测，并在训练期间与真实标签进行比较。

### 3.1 评论文本节点表征

类似于TextGCN(Yao et al., 2019)，论文中矩阵$\mathrm{X} = \mathrm{I}_{\mathrm{n_{doc}+n_{word}}}$被用作初始节点特征，其中$\mathrm{n_{doc}}$是文档节点的数量，$\mathrm{n_{word}}$是单词节点的数量（包括训练集和测试集），在本文中我们使用多语言预训练语言模型XLM-R来获得汉越双语评论文本的嵌入，并将它们作为异构图中评论文本节点的输入表示。

$$\mathrm{X} = \begin{pmatrix} X_{S_c} \\ X_{S_v} \\ 0 \\ 0 \end{pmatrix}_{(n_{S_c}+n_{S_v}+n_{K_c}+n_{K_v})\times \mathrm{d}} \tag{1}$$

其中，$n_{S_c}$、$n_{S_v}$、$n_{K_c}$、$n_{K_v}$分别表示中文评论文本数量、越南语评论文本数量、中文关键词数量和越南语关键词数量，中文评论文本节点和越南语评论文本节点嵌入由$\mathrm{X_{S_c}} \in \mathbb{R}^{n_{S_c}\times d}$和$\mathrm{X_{S_v}} \in \mathbb{R}^{n_{S_v}\times d}$表示，其中d是文本嵌入的维度。由于不考虑关键词节点的特征表示，因此将中文关键词和越南语关键词的嵌入置为0。

### 3.2 多语言异构图的构建

由于越南语标注数据资源稀缺，且以往所提出的表示学习模型仅学习到单语语料中的文本信息而忽略了同一事件下多语言观点对象之间的对齐关系。本文将汉越语料库中的各种实体和关系整合到一个异构图中，设计一种新的多语言表示学习模型，将语义信息和拓扑信息封装到一个低维联合嵌入的观点对象识别任务中，通过构建一个包含关键词节点和评论文本节点的异构图，节点数$n = (n_{S_c} + n_{S_v} + n_{K_c} + n_{K_v})$是中文和越南语评论文本数量和双语评论中关键词数量的总和，表1包含了异构图中节点和边的详细信息。

| No. | 节点 | 描述 |
|---|---|---|
| 1 | $S_c$ | 中文评论句 |
| 2 | $S_v$ | 越南语评论句 |
| 3 | $K_c$ | 中文关键词 |
| 4 | $K_v$ | 越南语关键词 |
| No. | 边 | 描述 |
| 1 | $S_c \leftrightarrow K_c$ | 中文评论句中包含中文关键词 |
| 2 | $S_v \leftrightarrow K_v$ | 越南语评论句中包含越南语关键词 |
| 3 | $K_c \leftrightarrow K_v$ | 中文关键词与越南语关键词词典 |
| 4 | $S_c \leftrightarrow S_v$ | 中文评论句与越南语评论句有较高的语义相似度 |

Table 1: 汉越多语言异构图的节点及边关系

使用汉越社交媒体评论文本数据集中的评论句和其中的关键词作为节点构建汉越多语言异构图，其中关键词之间、评论句和关键词以及评论句之间均有不同的关系种类，主要包括关键词之间的词共现和词对齐关系，评论句和关键词的词频关系，评论句之间的语义相似度关系。

**关键词之间的词共现和词对齐关系：**

为了更好的利用单语关键词的共现信息，通过基于词共现关系构建关键词节点之间的边，对语料库中所有的评论句使用一个固定大小的滑动窗口来收集词的共现信息，分别在中文和越南语评论文本上使用点互信息（PMI）计算两个关键词节点之间的权重，单语关键词对$\{i, j\}$的PMI值计算公式为：

$$PMI(i,j) = \log \frac{p(i,j)}{p(i)p(j)} \tag{2}$$

$$p(i,j) = \frac{\#W(i,j)}{\#W} \tag{3}$$

$$p(i) = \frac{\#W(i)}{\#W} \tag{4}$$

其中$\#W(i)$表示滑动窗口中包含关键词$i$的数量，$\#W(i,j)$是指滑动窗口中同时包含关键词$i$和$j$的数量，$\#W$是语料库中所有滑动窗口的数量。当PMI值为正时表示两个词之间的语

义相关性较高，而PMI值为负时表示两个词之间的语义相关性很少或没有，只在PMI值为正的关键词对之间添加边。

考虑挖掘汉越双语关键词之间的关系，基于双语词对齐构建关键词节点之间的边，对于汉越多语言观点对象识别的研究中，汉越双语关键词对相较于其他词对对模型预测性能产生的影响更大，利用汉越双语种子词典，匹配语义相似的双语关键词作为词节点并添加对齐的边关系，根据匹配出的双语关键词对进行多语言词级对齐和聚合，从而将两种语言的词级关系融入图结构中。

**评论句和关键词的词频关系：**

基于关键词在评论文本中出现的次数构建关键词与评论句之间的边，使用TF-IDF计算词频，其中TF是单词在评论句中出现的次数，IDF指的是由包含该单词的句子数量的对数缩放的逆分数，在评论句与关键词之间添加边并将计算的TF-IDF值作为边的权重。

**评论句之间的语义相似度关系：**

为了在评论句之间添加更直接的连接，使汉越两种语言的评论句可以更好的进行同一嵌入空间下的迁移学习，通过多语言预训练语言模型XLM-R得到汉越两种语言评论句的嵌入向量$(A_i, B_j)$，同时利用余弦相似度计算两个嵌入向量之间的相似性。

$$\cos\theta = \frac{A_i \cdot B_j}{|A_i| \times |B_j|} \tag{5}$$

其中$\mathrm{A_i} \in \mathrm{X_{S_c}}$表示第$i$条中文评论文本嵌入向量，$\mathrm{B_j} \in \mathrm{X_{S_v}}$表示第$j$条越南语评论文本嵌入向量。当余弦值越接近1表示两个向量的夹角越接近0度，也就是两个向量越相似，设置超参数Q作为阈值，找到余弦相似度最大的Q个评论文本添加边关系。

### 3.3 基于图卷积网络的节点特征学习

在构建多语言异构图后，将不同关系类别的子图进行融合，嵌入到一个简单的二层图卷积网络中。图卷积网络是一种多层神经网络，可以根据节点的领域属性引入节点的嵌入向量。GCN可以通过一层卷积来捕获关于近邻节点的信息，当堆叠多个GCN层时，图上更多的信息就会被整合起来。两层GCN可以允许信息在最多两步长的节点之间传递信息，对于一层GCN，新的$s$维节点特征矩阵$L^{(1)} \in \mathbb{R}^{n \times s}$为：

$$L^{(1)} = \rho\left(\tilde{A}XW_0\right) \tag{6}$$

其中$\tilde{A} = \widetilde{D}^{-1/2}A\widetilde{D}^{-1/2}$表示标准化对称邻接矩阵，$W_0 \in R^{d \times s}$表示权重矩阵。$\rho$是激活函数，本文使用的是RELU。通过叠加多个GCN层来学习合并更高阶的领域信息，学习更深层的节点特征。可以表示为：

$$L^{(j+1)} = \rho\left(\tilde{A}L^{(j)}W_j\right) \tag{7}$$

其中$j$表示层数，而$L^{(0)}$表示原始邻接矩阵。

### 3.4 评论文本观点对象类别预测

观点对象识别过程是判断当前节点属于哪一类别，属于分类过程。在图神经网络的第二层将评论文本嵌入维度映射成与类别标签相同的维度大小，然后送入到分类器中：

$$Z = \mathrm{softmax}\left(\tilde{A}\,\mathrm{Relu}\left(\tilde{A}XW_0\right)W_1\right) \tag{8}$$

其中$\mathrm{softmax} = \frac{1}{z}\exp(x_i)$，而$z = \sum_i \exp(x_i)$。

模型的损失函数使用交叉熵损失：

$$L = -\sum_{d \in y_D}\sum_{f=1}^{F} Y_{df} \ln Z_{df}。 \tag{9}$$

其中$y_D$是具有标签的评论索引集，$F$表示输出特征的维度，与类别数量相同，$Y$是标签矩阵。

# 4 实验设置

## 4.1 数据集

为了证明实验的有效性，本文参考Conneau et al. (2018)构建的XNLI多语言文本数据集格式，构建了汉越多语言观点对象识别数据集。利用网络爬虫技术在Twitter和新浪微博上爬取“新冠疫情”和“亚裔歧视”相关评论作为实验数据，通过语种识别方法清除非汉越数据，利用emoji数据包和正则表达式去除文本中的表情、符号以及超链接等，再进行数据筛查和整理完成数据清洗，对数据集按照6:2:2的比例划分训练语料、验证语料和测试语料，汉越观点对象数据集的具体划分信息和观点对象类别如表2，3所示:

| 种类 | 语种 | 训练语料 | 验证语料 | 测试语料 |
|---|---|---|---|---|
| 新冠疫情 | 中文 | 3000 | 1000 | 1000 |
| | 越南语 | 2000 | 600 | 600 |
| 亚裔歧视 | 中文 | 3000 | 1000 | 1000 |
| | 越南语 | 2000 | 600 | 600 |

Table 2: 汉越观点对象识别数据集（单位：条）

| 种类 | 新冠疫情 | 亚裔歧视 |
|---|---|---|
| 观点对象类别 | 新冠病毒 | 游行 |
| | 疫苗接种 | 亚裔 |
| | 疫情防控 | 种族歧视 |
| | 其它 | 其它 |

Table 3: 汉越评论文本观点对象类别

## 4.2 评价指标

与其他分类任务类似，本文实验使用测试数据集上准确度Acc（Accuracy）、精确度P（Precision）、召回率R（Recall）和F1值的结果作为评价指标，从而衡量模型的性能。公式如下:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

$$P = \frac{TP}{TP + FP} \tag{11}$$

$$R = \frac{TP}{TP + FN} \tag{12}$$

$$F1 = \frac{2PR}{(P + R)} \tag{13}$$

其中TP表示正类被正确预测，FP表示负类被错误预测，FN表示正类被错误预测，TN表示负类被正确预测。

## 4.3 实验参数设置

利用Adam优化器对图卷积网络和分类器进行联合优化，实验使用多语言预训练模型mBert得到汉越双语评论文本的特征表示，向量维度为768，对每个评论句取相似度最高的Q个评论句，使用dropout防止过拟合。设置模型的最大训练批次为100个，设置early stopping在10个批次后当连续10次Epoch（或者更多次）没达到最佳精度时则模型训练终止，在验证集上选择最佳模型，所有实验都在单个GeForce RTX 3090 GPU上进行，具体信息如下表4所示。

| 参数 | 值 |
|---|---|
| dropout | 0.5 |
| 学习率 | 0.0005 |
| 最大轮次 | 100 |
| 滑动窗口大小 | 20 |
| 评论文本相似度阈值Q | 3 |
| GCN层数 | 2 |
| GCN隐藏层维度 | 200 |

Table 4: 参数设置

### 4.4 基线模型

- MT+LM(Zhai and Lafferty, 2017)：将训练的越南语评论翻译为中文评论，利用预训练语言模型对评论句进行表征并训练观点对象分类器，最终在测试数据上实现观点对象识别。
- TF-IDF+LR(Yoo and Yang, 2015)：具有术语频率和反向文档频率加权的词袋模型加监督学习中经典的分类方法，以线性回归为理论支持，通过Sigmoid函数引入了非线性因素，解决分类任务。使用源语言中文训练的基线模型，并仅依靠双语词嵌入对目标进行分类。
- CNN(Kim, 2014)：采用TextCNN模型，使用源语言中文训练的基准模型，并仅依靠双语词嵌入对目标进行分类，设置卷积核大小为$\{3, 4, 5\}$。
- Node2vec(Grover and Leskovec, 2016)：Node2vec通过网络中的二阶随机游走来学习图的嵌入，通过在验证集上对$p, q \in \{0.25, 0.5, 1, 2, 4\}$进行网格搜索，为实验选择最佳的参数设置。
- MT+TextGCN(Yao et al., 2019)：将训练后的中文评论翻译为越南语评论，利用翻译后的文本进行异构图构建，并利用TextGCN对节点特征进行学习。
- CLHG(Wang et al., 2021)：使用基于异构图的图卷积网络，通过机器翻译对不同语言的文档进行翻译，文档和词之间存在的不同关系创建异构图结构。
- Bert+GCN(She et al., 2022)：使用图卷积网络获得评论文本的句法结构信息，多语言预训练语言模型获取文本的上下文信息，通过动态融合门对两个信息进行融合得到融合向量，对融合向量的文本进行识别分类。

## 5 实验结果分析

### 5.1 基线模型实验对比结果

表5列出了本文模型与基线模型在“新冠疫情”和“亚裔歧视”两个数据集上的实验对比结果。从实验结果可以看出，本文模型与其他基准模型相比有较大的优势，具体分析如下：

（1）将本文模型与MT+LM进行对比，以“新冠疫情”数据集为例，本文提出模型的F1值提升了25.51个百分点，分析原因在于翻译得到的标注语料含有大量噪声，同时只使用双语词嵌入的方法尚不具备捕获中文和越南语评论中观点对象关联信息的能力。

（2）对比本文模型与TF-IDF+LR基线时，本文模型的性能同样高于该基线的性能。推测原因是两种语言具有完全不同的词汇表，在多语言数据集上基于词袋模型捕获到的双语特征差距大，而MT+LM的Accuracy相比较TF-IDF+LR要高出1-6个百分点，说明机器翻译能够起到弥补语义鸿沟的问题。

（3）对比本文模型与CNN基线模型时，以”新冠疫情”数据集为例，本文模型的Accuracy和F1值分别高出23个和12个百分点。而CNN模型性能要比MT+LM模型性能好，说明利用CNN能够编码出更好的评论特征，同时也验证了仅使用嵌入向量无法完成观点对象关联信息捕获的问题。

（4）分析本文模型与Node2vec、MT+TextGCN的结果，Node2vec模型是基于同构网络设计，比直接对评论进行特征编码的性能有所提升，这一观察结果证实了异构信息能够提高模型

| 数据集 | 方法 | Acc | P | R | F1 |
|---|---|---|---|---|---|
| 新冠疫情 | TF-IDF+LR | 0.5358 | 0.4455 | 0.4321 | 0.4796 |
| | MT+LM | 0.5900 | 0.42204 | 0.4241 | 0.4512 |
| | CNN | 0.7286 | 0.5429 | 0.5333 | 0.5810 |
| | Node2vec | 0.7600 | 0.6514 | 0.6243 | 0.6084 |
| | MT+TextGCN | 0.7700 | 0.6481 | 0.6067 | 0.6264 |
| | CLHG | 0.6920 | 0.6537 | 0.6537 | 0.6537 |
| | Bert+GCN | 0.9250 | 0.6743 | 0.6538 | 0.6639 |
| | 本文模型 | **0.9625** | **0.7028** | **0.7098** | **0.7063** |
| 亚裔歧视 | TF-IDF+LR | 0.5080 | 0.4167 | 0.4255 | 0.4426 |
| | MT+LM | 0.5150 | 0.4421 | 0.3608 | 0.4480 |
| | CNN | 0.6900 | 0.5725 | 0.5250 | 0.5204 |
| | Node2vec | 0.6975 | 0.6695 | 0.6284 | 0.6457 |
| | MT+TextGCN | 0.7100 | 0.6593 | 0.6546 | 0.6569 |
| | CLHG | 0.7220 | 0.7278 | 0.7278 | 0.7278 |
| | Bert+GCN | 0.9400 | 0.8446 | 0.8552 | 0.8499 |
| | 本文模型 | **0.9625** | **0.9607** | **0.8787** | **0.9179** |

Table 5: 汉越多语言事件观点对象识别方法性能对比

的表示学习能力。而MT+TextGCN模型的Accuracy和F1值相较于Node2vec普遍有1-3个百分点的提升，分析原因认为在进行节点特征学习的过程中，TextGCN将节点的新特征计算为节点自身及其二阶邻居节点的加权平均值，使得评论节点的标签信息能够进一步传递到相邻的其他评论节点和词节点中。

（5）对比本文模型与CLHG模型时，本文提出模型的Accuracy和F1值均高于CLHG模型，分析原因认为相比利用机器翻译缩小语言差异，利用词节点收集全面的评论标签信息，并且利用语义相似度计算捕获图中关联信息作为异构图中的关键路径，从而使标签信息可以传播到整个图中。

（6）分析本文模型与Bert+GCN的结果，以“新冠疫情”数据集为例，本文模型的Accuracy和F1值分别高出3.75个和4.24个百分点，多语言预训练语言模型可以学习到评论文本上下文的语义特征信息，有利于模型性能的提升，同时证实了图卷积网络能够学习邻居节点的特征信息，提高模型的表示学习能力。

### 5.2 不同多语言预训练语言模型的实验结果

为了验证不同的多语言预训练语言模型对本文模型方法的影响，本文分别使用mBert、XLM和XLM-R对数据集中的评论文本节点进行表征，词嵌入维度分别为786、1280和786，其它所有参数设置均相同，实验结果如表6所示。

| 数据集 | 多语言预训练语言模型 | Acc | R | F1 |
|---|---|---|---|---|
| 新冠疫情 | mBert | 0.9318 | 0.6741 | 0.6638 |
| | XLM | 0.9475 | 0.6891 | 0.6893 |
| | XLM-R（本文） | **0.9625** | **0.7098** | **0.7063** |
| 亚裔歧视 | mBert | 0.9336 | 0.8306 | 0.8714 |
| | XLM | 0.9575 | 0.8505 | 0.8982 |
| | XLM-R（本文） | **0.9625** | **0.8787** | **0.9179** |

Table 6: 不同多语言预训练语言模型对实验结果的影响

观察表6可以发现，在使用不同的多语言预训练语言模型对节点进行表征，所有参数设置相同时，选择多语言预训练模型XLM-R做表征时模型效果最好。

### 5.3 图卷积层数设定对实验结果的影响

考虑到在图卷积学习的过程，图卷积层数的设定对聚合邻居节点信息程度有影响。本节针对图卷积层数在“新冠疫情”和“亚裔歧视”两个数据集上进行实验分析，实验结果如下图3所示：

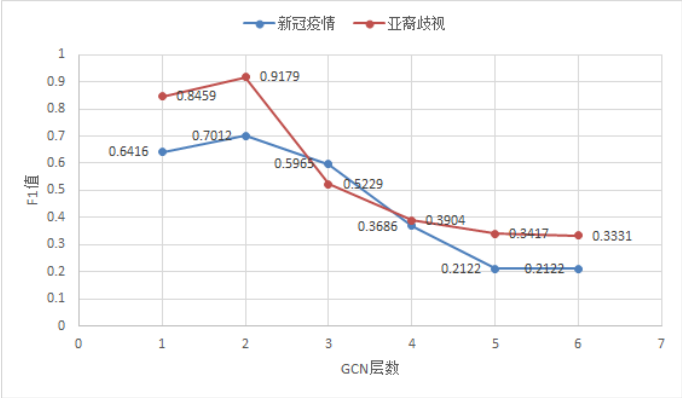


Figure 3: GCN层数设定对模型结果的影响

图3显示了在两个数据集上不同的卷积层数下模型F1值的结果，可以观察到模型的F1值首先随着卷积层数的增大而增加，当卷积层数为2时，模型的性能达到最佳，这表明卷积层数不足2时，卷积网络对信息聚合的能力不足，模型性能较低，当卷积层数超过2层后，随着层数的增加，模型性能有所下降并趋于稳定，因此本文提出的模型中将卷积层数的大小设定为2。

### 5.4 不同图结构对实验结果的影响

在本文中，我们通过使用汉越社交媒体评论文本数据集中的评论句和其中关键词作为节点构建汉越多语言异构图，构建了四种不同关系种类的子图，其中单语言图包括关键词之间的词共现关系以及评论文本和关键词的词频关系，多语言图包括词对齐关系和评论文本之间的语义相似度关系。本节根据不同构图方式在“新冠疫情”和“亚裔歧视”两个数据集上进行实验分析，实验性能的对比结果如下表7所示：

| 数据集 | 构图方法 | Acc | R | F1 |
|---|---|---|---|---|
| 新冠疫情 | 单语言图 | 0.9250 | 0.6538 | 0.6639 |
| | 多语言图 | 0.9425 | 0.6993 | 0.6862 |
| | 本文 | **0.9625** | **0.7098** | **0.7063** |
| 亚裔歧视 | 单语言图 | 0.9400 | 0.8552 | 0.8499 |
| | 多语言图 | 0.9475 | 0.8501 | 0.8677 |
| | 本文 | **0.9625** | **0.8787** | **0.9179** |

Table 7: 不同图结构的性能对比

观察表7可以发现，在“新冠疫情”和“亚裔歧视”上构建多语言图比构建单语言图的F1值高，证明了进行跨语言的有效性，而本文方法是将单语言的两种子图和多语言的两种子图进行融合，所取得的F1值均高于单独构建子图的效果，实验结果表明添加不同类型边的关系会取得更好的性能。

## 6 结论

本文提出了一种融合汉越关联关系的多语言事件观点对象识别方法，利用关联事件下的汉越社交媒体评论文本数据作为模型训练语料，结合汉越评论文本之间以及关键词之间的关联关系构建多语言异构图，随后利用图卷积网络对该图进行建模，从而聚合邻居节点信息并捕获高阶领域信息，利用该方法能够识别出汉越双语评论文本中的观点对象，实验结果证明了本文所提方法的有效性。下一阶段的工作我们将重点研究如何融入观点对象信息对汉越评论文本进行情感倾向性的分析。

## 参考文献

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. *arXiv preprint arXiv:1809.05053*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Ying Ding, Jianfei Yu, and Jing Jiang. 2017. Recurrent neural networks with auxiliary labels for cross-domain opinion target extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. corr abs/1408.5882. *arXiv preprint arXiv:1408.5882*.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Fangtao Li, Sinno Jialin Pan, Ou Jin, Qiang Yang, and Xiaoyan Zhu. 2012. Cross-domain co-extraction of sentiment and topic lexicons. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 410–419.

Xin Li, Lidong Bing, Piji Li, Wai Lam, and Zhimou Yang. 2018. Aspect term extraction with history attention and selective transformation. *arXiv preprint arXiv:1805.00760*.

Zheng Li, Mukul Kumar, William Headden, Bing Yin, Ying Wei, Yu Zhang, and Qiang Yang. 2020. Learn to cross-lingual transfer with meta graph learning across heterogeneous languages. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 2290–2301.

Samaneh Moghaddam and Martin Ester. 2011. Ilda: interdependent lda model for learning latent aspects and their ratings from online product reviews. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 665–674.

Huy Tien Nguyen and Minh Le Nguyen. 2018. Multilingual opinion mining on youtube–a convolutional n-gram bilstm word embedding. *Information Processing & Management*, 54(3):451–462.

Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. 2011. Opinion word expansion and target extraction through double propagation. *Computational linguistics*, 37(1):9–27.

Xiangrong She, Jianpeng Chen, and Gang Chen. 2022. Joint learning with bert-gcn and multi-attention for event text classification and event assignment. *IEEE Access*, 10:27031–27040.

Ivan Titov and Ryan McDonald. 2008. Modeling online reviews with multi-grain topic models. In *Proceedings of the 17th international conference on World Wide Web*, pages 111–120.

Ziyun Wang, Xuan Liu, Peiji Yang, Shixing Liu, and Zhisheng Wang. 2021. Cross-lingual text classification with heterogeneous graph neural network. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 612–620.

Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. Graph convolutional networks for text classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 7370–7377.

Jong-Yeol Yoo and Dongmin Yang. 2015. Classification scheme of unstructured text document using tf-idf and naive bayes classifier. *Advanced Science and Technology Letters*, 111(50):263–266.

Chengxiang Zhai and John Lafferty. 2017. A study of smoothing methods for language models applied to ad hoc information retrieval. In *ACM SIGIR Forum*, volume 51, pages 268–276. ACM New York, NY, USA.

倪茂树and 林鸿飞. 2007. 基于关联规则和极性分析的商品评论挖掘. In 第三届全国信息检索与内容安全学术会议, volume 635, page 642.