

基于文本语义和社交行为信息融合的讽刺检测方法

付朝阳^{1,2,3}, 陈致凯^{1,2,3}, 潘理^{1,2,3}

(1. 上海交通大学网络安全技术研究院, 上海 200240; 2. 信息安全综合管理技术研究上海市重点实验室, 上海 200240;
3. 上海交通大学张江高等研究院, 上海 200240)

摘要: 讽刺是一种复杂的隐式情感, 讽刺检测是文本情感分析领域的重要研究问题, 对于社交网络舆情分析有极强的现实意义。讽刺文本的表层语义和用户的真实情感往往相反, 基于表层语义的文本情感检测通常会得到错误的分类结果。日常交流中的讽刺往往借助语调、神态等非文本信息进行表达, 因此单纯基于文本语义的讽刺检测方法无法利用非文本信息, 极大地制约了检测效果。为了充分利用文本语义与用户社交行为信息, 提升讽刺检测效果, 提出了一种基于文本语义和社交行为信息融合的讽刺检测方法。该方法构造了包含用户、文本、情感词的异质信息网络, 并设计了一种用于异质信息图表征向量计算的图神经网络模型。该模型使用双重注意力机制提取社交行为信息, 通过情感子图挖掘文本深层语义, 最终得到融合文本语义和社交行为信息的融合特征向量。融合特征向量可以用于训练分类器, 进而完成社交网络文本讽刺检测。在推特真实数据集上的充分实验表明, 所提方法的分类效果优于现有的讽刺文本检测方法。

关键词: 讽刺检测; 图神经网络; 异质信息融合; 隐式情感分析

中图分类号: TP183

文献标志码: A

DOI: 10.11959/j.issn.2096-109x.2023059

Sarcasm detection method based on fusion of text semantics and social behavior information

FU Zhaoyang^{1,2,3}, CHEN Zhikai^{1,2,3}, PAN Li^{1,2,3}

1. Institute of Cyber Science and Technology, Shanghai Jiao Tong University, Shanghai 200240, China
2. Shanghai Municipal Key Lab of Integrated Management Technology for Information Security, Shanghai 200240, China
3. Zhang jiang Institute for Advanced Study, Shanghai Jiao Tong University, Shanghai 200240, China

Abstract: Sarcasm is a complex implicit emotion that poses a significant challenge in sentiment analysis, particularly in social network sentiment analysis. Effective sarcasm detection holds immense practical significance in the analysis of network public opinion. The contradictory nature of sarcastic texts, which exhibit implicit semantics opposite to the real emotions of users, often leads to misclassification by traditional sentiment analysis

收稿日期: 2022-09-30; 修回日期: 2023-02-17

通信作者: 潘理, panli@sjtu.edu.cn

基金项目: 国家自然科学基金(62172278)

Foundation Item: The National Natural Science Foundation of China (62172278)

引用格式: 付朝阳, 陈致凯, 潘理. 基于文本语义和社交行为信息融合的讽刺检测方法[J]. 网络与信息安全学报, 2023, 9(4): 134-143.

Citation Format: FU Z Y, CHEN Z K, PAN L, et al. Sarcasm detection method based on fusion of text semantics and social behavior information [J]. Chinese Journal of Network and Information Security, 2023, 9(4): 134-143.

methods. Moreover, sarcasm in daily communication is often conveyed through non-textual cues such as intonation and demeanor. Consequently, sarcasm detection methods solely relying on text semantics fail to incorporate non-textual information, thereby limiting their effectiveness. To leverage the power of text semantics and social behavior information, a sarcasm text detection method based on heterogeneous graph information fusion was proposed. The approach involved the construction of a heterogeneous information network encompassing users, texts, and emotional words. A graph neural network model was then designed to handle the representations of the heterogeneous graph. The model employed a dual-channel attention mechanism to extract social behavior information, captured the deep semantics of text through emotional subgraphs, and ultimately combined text semantics and social behavior information. Extensive experiments conducted on the Twitter dataset demonstrate the superiority of the proposed method over existing approaches for sarcasm text detection and classification.

Keywords: sarcasm detection, graph neural network, heterogeneous information fusion, implicit sentiment analysis

0 引言

讽刺是一种复杂的负面隐式情感。在讽刺文本中,字面内容表达的情感和用户真实的情感倾向完全相反,传统的文本情感分析方法往往会得到错误的结果。在社交网络中广泛存在含有讽刺的文本内容,这些文本对传统的情感分析和网络舆情态势感知工具构成了严峻挑战,因而针对社交网络讽刺文本的检测成为近年来情感分析领域的研究热点^[1-2]。

在显式讽刺文本中,文本内部的情感不协调是讽刺表达的重要特征。但是一些隐式讽刺文本并不含有明显的情感冲突,单独使用文本语义无法准确判定。在面对面交流中,用户更多地会使用语音语调、表情神态等非文本信息表达讽刺情感。社交媒体中的文本缺少这些重要的非文本信息,这给文本讽刺检测任务增加了挑战^[3]。但是,社交网络中的文本也天然地包含了社交行为信息,如其他用户在当前文本下的留言互动、当前文本内容发布者的文本表达习惯等,这些信息对于讽刺检测有着重要的参考价值^[4]。如何充分利用社交网络媒体信息,挖掘文本语义和社交行为信息的融合特征是社交网络讽刺检测的关键。

本文提出了一种基于文本语义和社交行为信息融合的讽刺检测方法:首先,基于异质信息网络(HIN, heterogeneous information network)建构了一种包含用户、文本、情感关键词等多种节点的社交异质信息网络(UTE-HIN, User, Text, Emotion-HIN)。然后提出社交网络信息融合图注意力网络(SNIF-GAT, social network information fusion graph attention network)模型,模型可以对

社交异质信息网络中的文本、用户等节点进行表征计算。SNIF-GAT是一种异质图注意力网络模型,模型输出的特征向量融合了文本语义和社交行为等信息。融合特征向量可以应用于下游节点分类任务,从而完成社交网络文本讽刺检测的任务。实验结果表明,本文方法在基准数据集上取得了比传统文本讽刺检测方法更好的表现。

1 相关工作

讽刺是一种在日常生活中常见的复杂情绪表达,随着社交媒体在近年来的广泛使用,社交媒体上也出现了大量讽刺文本,讽刺与非讽刺文本的示例如图1所示。早期的研究工作从语言学出发,围绕显式讽刺文本中的情感冲突构造结构特征进行检测^[5],一些基于修辞学特征工程的机器学习方法也取得了良好的成绩^[6-7]。随着深度学习的发展,CNN^[8](convolutional neural network)、LSTM^[9](long short term memory)等神经网络方法被应用于讽刺文本检测。2018年以来,以BERT(bidirectional encoder representations from transformers)为代表的预训练模型在自然语言处理任务中得到广泛应用,使用BERT的讽刺文本检测方法显著提升了传统神经网络模型的效果^[10-11]。

为了提升检测效果,一方面需要挖掘深层语义,另一方面需要引入非文本附加信息用于辅助检测。在深层语义挖掘上,Tay等^[12]提出了MIARN(multi-dimensional intra-attention recurrent network)模型,模型使用内部注意力机制学习文本内部的情感冲突,同时使用LSTM生成全文表征向量,通过情感冲突表征向量与全文表征向量

的混合使用进行讽刺检测。在附加信息引入上, Hazarika^[13]提出了使用用户历史发帖数据的 CASCADE 模型, 模型根据用户历史发帖数据、当前帖子的回复内容生成用户表征向量, 同时使用 CNN 提取文本表征向量。CASCADE 综合使用用户表征向量和文本表征向量进行讽刺检测, 相比传统神经网络讽刺检测模型有了显著的性能提升。

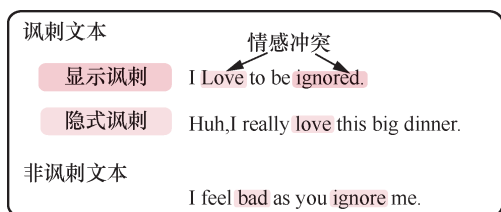


图1 讽刺与非讽刺文本的示例

Figure 1 Examples of sarcastic and non-sarcastic text

自2016年 Kipf 等^[14]提出图卷积网络(GCN, graph convolution network)以来, 基于 GCN 的各种模型蓬勃发展。由于图卷积网络在处理图结构数据上的优势, 一些传统任务在引入图的方法后取得了良好效果^[15]。GCN 在讽刺检测中也得到了广泛使用, Lou^[16]提出了基于 GCN 的 ADGCN (affective dependency graph convolutional network), 模型使用情感依赖图和语法依赖图来挖掘文本深层语义。Lin 等^[17]提出了基于图注意力网络(GAT, graph attention network)的 ARGAT (affection enhanced relational graph attention network), 通过注意力机制增强语义表达能力。为了引入社交行为信息, Plepi 等^[18]提出融合用户特征和文本特征的 GAT, 该方法构造了包含文本和用户历史数据两种类型节点的同质图, 通过注意力机制计算用户特征向量, 同时使用 BERT 根据文本内容计算文本特征向量。在包含社交信息的数据集 SPIRS^[19]上的实验中, GAT 虽然缺乏深层语义挖掘能力, 但附加了社交行为信息, 性能相比基于文本语义的模型取得了显著提升。文本语义和社交网络都包含图结构, 图神经网络在文本讽刺检测中既可以挖掘深层语义, 又可以提取社交行为信息。因此, 本文提出了一种基于文本语义和社交行为信息融合的讽刺检测方法, 针对社交网络特点建构异质信息网络, 并提出 SNIF-GAT 模型, 在 UTE-HIN 中计算融合特征向量。融合特征向量用于训练分类器, 完成训练的分类器可以

实现讽刺检测的分类任务。

2 异质信息网络构建与节点初始化

异质信息网络可表示为有向图结构 $G=(V, E, \Phi, \Psi)$, 其中, V 表示节点集合, E 表示关联节点间连接形成的边。每个异构网都包含映射关系 $\Phi: v \rightarrow A$ 以及 $\Psi: e \rightarrow R$, 前者表示节点 $v \in V$ 与节点类型之间的映射, 后者则表示边 $e \in E$ 与边属性间的映射, 而 A, R 则分别表示节点属性与边属性的集合。不同属性的节点与边彼此连接构成异质信息网络。

为了融合社交网络中的文本语义与社交行为信息, 本文设计了一种包含用户、文本、情感关键词的社交异质信息网络, 网络结构如图2所示。其中, 节点类型集合 $A=\{U, T, K\}$, 分别表示用户、文本、情感关键词3种节点类型。边属性集合 $R=\{U-U, U-T, T-T, T-K\}$, 分别表示不同类型节点间的连接关系。

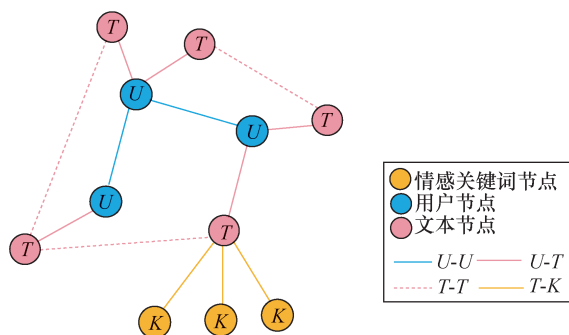


图2 社交异质信息网络结构

Figure 2 The architecture of the UTE-HIN

在节点类型集合 A 中, 对于用户类型的节点, 网络中任意一个用户节点用 u_i 表示。用户节点 u_i 的属性为用户特征向量, 节点在网络中通过聚合邻居节点的信息来更新特征向量, 特征向量主要包含用户的社交行为信息。对于文本类型节点, 网络中任意一个文本节点用 t_i 表示。文本节点 t_i 的属性为文本特征向量, 节点在情感关键词构成的情感子图上更新特征向量, 特征向量主要包含文本语义信息。对于情感关键词类型节点, 网络中任意一个情感关键词节点用 k_i 表示。情感关键词节点 k_i 的属性不是特征向量, 而是情感词典中的一个情感词, 情感词典通过对帖子文本中的情感词进行词频分析得到。在边属性集合 R 中, 当

两个用户存在评论等互动行为, 两个用户节点间建立 $U-U$ 类型的连接。当某用户发布新帖子, 用户节点和帖子对应的文本节点间建立 $U-T$ 类型的连接。当某用户评论其他用户的帖子, 两个帖子对应的文本节点间建立 $T-T$ 类型的连接。如果两个用户节点间尚未建立 $U-U$ 型连接, 则两个用户节点间同时建立 $U-U$ 类型的连接。文本节点对应的原始输入是一条帖子的文本内容, 如果帖子文本中包含某个情感关键词, 则文本节点和该情感关键词对应的情感关键词节点间建立 $T-K$ 类型的连接。通过情感关键词的包含关系, 某个文本节点可以和多个情感关键词节点构成一个情感子图。情感子图包含文本语义和情感依赖信息, 可以在子图上挖掘文本深层语义。

社交异质信息网络的输入包括用户 ID 和帖子文本两部分, 分别对应一个用户和文本节点, 根据用户发帖和评论关系完成 UTE-HIN 的初步构建。为了确定情感关键词的内容和数量, 首先对全体帖子文本进行情感词词频分析, 然后从 SenticNet^[20]提供的推特情感词典中匹配若干高频的情感词, 每个情感词对应一个情感关键词节点。根据文本和情感关键词的包含关系, 构建文本节点的情感子图。UTE-HIN 中的用户节点和文本节点采用不同的初始化方法。对于用户节点 u_i , 节点的原始输入为用户 ID, 首先根据用户 ID 从历史帖子数据库中获取该用户历史发布的帖子, 使用 user2vec^[21]根据用户历史发布的多条帖子内容生成特征向量, 该特征向量即用户节点 u_i 的初始化特征向量 h_u^0 。对于文本节点, 节点的原始输入为帖子文本, 首先使用 NLTK 工具^[22]对文本进行分词处理, 然后使用 BERT 生成每个单词的表征向量, 这些表征向量组合形成文本节点的初始化特征向量 h_t^0 。至此, UTE-HIN 的构建和节点特征向量的初始化完成。为了使后文的叙述清晰准确, 本文使用的各类符号及对应含义如表 1 所示。

3 基于社交网络信息融合模型的讽刺检测框架

3.1 问题定义

社交网络文本讽刺检测是一个二元分类问题, 主要目的是将社交网络中的文本内容分类为

讽刺文本和非讽刺文本。给定一组来自社交媒体的文本 $D = \{(t_1, u_1), (t_2, u_2), \dots, (t_n, u_n)\}$, 其中 t_i 是用户 u_i 在社交网络中发布的一段文本内容, n 表示待检测的文本的数量。通过讽刺分类方法 $f: D \rightarrow Y$, 将文本 t_i 分类为预定义类别 $Y = \{0, 1\}$ 。其中, 0 表示非讽刺文本, 1 表示讽刺文本。

表 1 符号及对应含义
Table 1 Symbol and meanings

符号	含义
V	节点集合, $V = \{u_1, u_2, \dots, t_1, t_2, \dots, k_1, k_2, \dots\}$
A	节点类型集合, $A = \{U, T, K\}$
u_i	第 i 个用户 (U) 类型的节点
t_i	第 i 个文本 (T) 类型的节点
k_i	第 i 个情感关键词 (K) 类型的节点
R	边属性集合, $R = \{U-U, U-T, T-T, T-K\}$
x_i	一条帖子中第 i 个单词
S	由 $[x_1, x_2, \dots, x_n]$ 组成的文本的词向量矩阵
h_u^0	用户节点 u_i 的初始化特征向量
h_t^0	文本节点 t_i 的初始化特征向量
h_u	用户节点 u_i 完成更新的特征向量
h_t	文本节点 t_i 完成更新的特征向量
h_f	融合特征向量

3.2 讽刺检测框架

本文提出的讽刺检测方法框架包含社交异质信息网络 (UTE-HIN)、SNIF-GAT 模型、分类检测器三大模块, 如图 3 所示。首先, 基于 UTE-HIN 进行异质信息网络构建并完成节点的初始化。其次, 使用 SNIF-GAT 模型在异质信息网络中对用户节点和文本节点的特征向量进行计算和更新, 拼接文本和用户的特征向量得到融合特征向量。最后, 将融合特征向量输入基于多层感知机 (MLP, multi-layer perceptron) 的分类网络, 分类器最终输出文本类别, 完成讽刺检测。

本文面向的讽刺检测为节点二元分类任务, 因此选择经典的 MLP 作为分类器。MLP 结构简单, 在多种分类任务中都有良好表现^[23]。本文使用的 MLP 包含两个隐藏层, 输出层使用 Softmax 对分布概率进行归一化, 得到二元分类结果的概率分布, 最终输出待检测的帖子文本是否为讽刺类型。

文本特征向量和用户特征向量的计算是本文的核心工作, 为了在异质信息网络中计算节点表

征,本文提出的 SNIF-GAT 模型在 GCN 的基础上引入了双重注意力机制和情感子图。

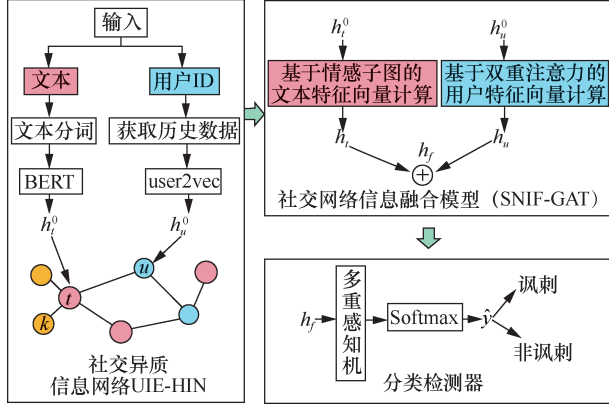


图3 基于 SNIF-GAT 的讽刺检测方法框架
Figure 3 The framework of the sarcasm detection method based on SNIF-GAT

3.3 基于情感子图的文本特征向量计算

为了使讽刺情绪被充分感知,显式讽刺文本通常包含情感冲突。例如,图1提供的显式讽刺文本“I love to be ignored”中的“love”和“ignored”即存在明显情感冲突。给定一个文本节点 t_i ,节点的文本经过分词后变成一个单词序列 $[x_1, x_2, \dots, x_n]$,其中 x_i 为序列中第 i 个单词, n 是文本中包含单词的总数。使用 BERT 生成每个单词 x_i 的词向量 $\mathbf{x}_i$,则词向量矩阵 $\mathbf{S}=[\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n]$ 可以表征当前文本的表层语义,该矩阵即文本节点的初始化特征向量 $\mathbf{h}_i^0$ 。本文使用情感邻接矩阵(AFF, affection adjacent matrix)提取并放大词向量矩阵 $\mathbf{S}$ 内部的情感冲突,实现对文本特征向量的更新。

为了构建情感子图,UTE-HIN 对总体文本的词频分析,从 SenticNet 提供的情感词典中匹配出若干高频情感词构成情感关键词节点,如图4所示,情感子图构建中的“love”“ignore”“cheers”

均为情感关键词节点。如果文本中包含某个情感关键词,则文本节点和相应的情感关键词节点建立连接,遍历单词序列后完成情感子图建构。情感子图的结构通过情感邻接矩阵 $\mathbf{AFF}$ 进行描述,矩阵中元素的计算方法如下。

$$\mathbf{AFF}_{i,j} = \begin{cases} 1 + |w(x_i) - w(x_j)|, & x_i, x_j \text{ 为不同的情感词} \\ 1, & i = j \\ 0, & \text{其他} \end{cases} \quad (1)$$

式中, $w(x_i)$ 表示单词 x_i 连接的情感关键词的权重。SenticNet 的情感词典包含权重,取值范围为 $[-1,1]$ 。若单词 x_i 包含负面情绪色彩,则权重值为负,反之则为正。若单词 x_i 和 x_j 存在情感冲突,则 $w(x_i)$ 和 $w(x_j)$ 的正负相反,如图4中的计算情感邻接矩阵所示,该词对在情感邻接矩阵中的权重值明显大于其他非情感关键词词对的权重值。如果 x_i 、 x_j 不全为情感关键词,则情感邻接矩阵中对应权重取值为0,降低了非情感关键词的噪声干扰。为了在文本特征向量中保留非情感关键词的表层语义,在情感子图中增加了自环,即 $\mathbf{AFF}$ 的对角线元素均为1。本文提出的讽刺框架使用了多层 SNIF-GAT,第 $l+1$ 层中文本节点的词向量矩阵 $\mathbf{S}^{l+1}$ 通过如下公式进行更新。

$$\mathbf{S}^{l+1} = \sigma(\widetilde{\mathbf{AFF}} \mathbf{S}^l \mathbf{w} + \mathbf{b}) \quad (2)$$

式中, $\mathbf{S}^l$ 是第 l 层 SNIF-GAT 更新后的词向量矩阵,第一层 SNIF-GAT 使用 BERT 生成的初始化文本特征向量 $\mathbf{h}_i^0$ 。 $\widetilde{\mathbf{AFF}}$ 是归一化的情感邻接矩阵, $\mathbf{w}$ 、 $\mathbf{b}$ 是第 $l+1$ 层网络需要训练的权重向量和偏置向量, σ 是网络使用的激活函数。

根据式(2)的更新法则,初始化词向量矩阵 $\mathbf{S}$

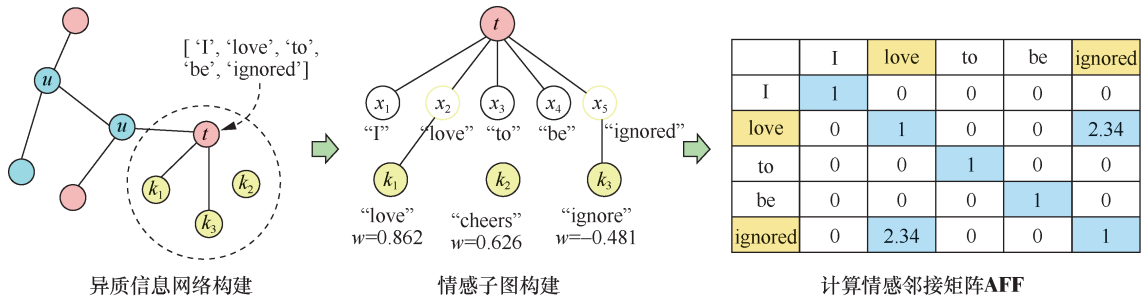


图4 情感子图与情感邻接矩阵
Figure 4 Affective subgraph and affective adjacency matrix

经过多层更新后变为 $\mathbf{S}' = [x'_1, x'_2, \dots, x'_n]$ ，本文使用注意力系数对 $\mathbf{S}'$ 中的单词向量加权求和，得到最终的文本特征向量 $\mathbf{h}_i$ 。本文的注意力系数受到 Zhang 等^[24]工作的启发，使用注意力系数 α_i 计算文本特征向量 $\mathbf{h}_i$ 的公式如下。

$$\mathbf{h}_i = \sum_{i=1}^n \alpha_i x'_i \quad (3)$$

$$\alpha_i = \frac{\exp(\beta_i)}{\sum_{k=1}^n \exp(\beta_k)} \quad (4)$$

$$\beta_i = \sum_{k=1}^n x_i^T x'_k \quad (5)$$

式中， $\mathbf{h}_i$ 是文本节点最终输出的特征向量， $\mathbf{S}' = [x'_1, x'_2, \dots, x'_n]$ 是最后一个隐藏层输出的特征向量矩阵， x_i^T 是词向量的转置。

3.4 基于双重注意力机制的用户特征向量计算

隐式讽刺文本通常不包含明显的情感冲突，需要结合用户的社交行为综合判定。例如，图 1 提供的隐式讽刺文本“Huh, I really love this big dinner”中没有显在的情感冲突，但结合用户的表达风格得知“love”为反语，该文本为讽刺文本。用户的历史发帖内容可以反映用户的表达习惯和文本风格^[25-26]，本文使用 user2vec 根据用户的历史数据计算用户节点的初始化特征向量 $\mathbf{h}_u^0$ 。为了在异质信息网络中聚合不同邻居文本和用户节点信息，更新用户特征向量，SNIF-GAT 在 HGAT^[27] 的启发下引入了基于边和节点的双重注意力机制，基于双重注意力机制的用户特征向量更新如图 5 所示。

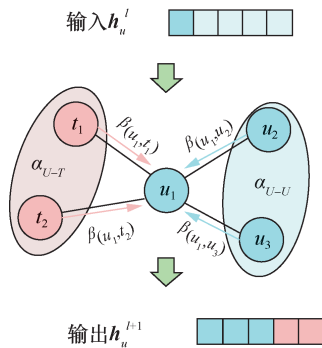


图5 基于双重注意力机制的用户特征向量更新
Figure 5 Update of user feature vectors with dual-channel attention mechanism

边注意力机制区分不同类型的边的重要程度，如图 5 所示，用户节点 u_i 通过 $U-U$ 与 $U-T$ 两种不同类型的边与其他节点连接，这两种边类型分别使用不同的边注意力系数 α_{U-U} 和 α_{U-T} 。对于边类型 $r \in R$ ，边注意力系数 α_r 的计算方法如下。首先对所有通过边类型 r 连接的邻居节点进行特征向量求和得到 $\mathbf{h}_r$ 。然后根据当前用户节点的特征向量 $\mathbf{h}_{u_i}$ 和 $\mathbf{h}_r$ 计算边类型 r 的注意力权重 $w_r = \sigma(a^T [h_{u_i} \parallel h_r])$ ，式中， a^T 是注意力向量的转置， $\parallel$ 表示向量的拼接运算， σ 是激活函数。最后使用 Softmax 对边属性集合 R 中的全部边类型的注意力权重进行归一化，得到边类型 r 的边注意力系数。

$$\alpha_r = \frac{\exp(w_r)}{\sum_{\tau \in R} (w_\tau)} \quad (6)$$

在通过相同类型边进行连接的节点中，节点注意力机制区分不同节点的重要程度。如图 5 所示，用户节点 u_i 和文本节点 t_1 、 t_2 间的连接类型相同，使用节点注意力系数 $\beta_{(u_i, t_1)}$ 、 $\beta_{(u_i, t_2)}$ 区分两个文本节点信息的重要程度。对于与用户节点 u_i 通过边类型 r 进行连接的邻居节点 n_j ，节点注意力系数 $\beta_{(u_i, n_j)}$ 的计算方法如下。首先计算边类型 r 的边注意力系数 α_r 。然后根据 u_i 和 n_j 的特征向量计算节点注意力权重 $w_{(u_i, n_j)} = \sigma(a^T \cdot \alpha_r [h_{u_i} \parallel h_{n_j}])$ ，式中， a^T 是注意力向量的转置。最后使用 Softmax 对邻居节点集合 N 中所有节点的节点注意力权重进行归一化，得到节点 u_i 和 n_j 的节点注意力系数。

$$\beta_{(u_i, n_j)} = \frac{\exp(w_{(u_i, n_j)})}{\sum_{v \in N} \exp(w_{(u_i, v)})} \quad (7)$$

在经典的多层 GCN 中，第 $l+1$ 层中节点的特征向量 $\mathbf{h}^{l+1}$ 通过如下传播规则进行更新。

$$\mathbf{h}^{l+1} = \sigma(\tilde{A} \mathbf{h}^l \mathbf{w} + \mathbf{b}) \quad (8)$$

式中， $\tilde{A}$ 是归一化的邻接矩阵， $\mathbf{h}^l$ 是第 l 层输出的特征向量， $\mathbf{w}$ 、 $\mathbf{b}$ 是第 $l+1$ 层网络需要训练的权重向量和偏置向量， σ 是网络使用的激活函数。SNIF-GAT 在 GCN 中引入双重注意力机制，通过

节点注意力系数为不同节点独立分配权重。将邻接矩阵对应位置的元素替换为节点注意力系数得到注意力邻接矩阵 (ATT, attention adjacent matrix)。ATT 不仅包含图的结构信息, 也可以区分不同节点和连接, 应用于异质信息网络的表征计算。在本文提出的讽刺检测框架中, 第 $l+1$ 层 SNIF-GAT 的用户特征向量 h_u^{l+1} 通过如下公式进行更新。

$$h_u^{l+1} = \sigma(\widetilde{\text{ATT}} h_u^l \mathbf{w} + \mathbf{b}) \quad (9)$$

式中, h_u^l 是第 l 层 SNIF-GAT 输出的用户特征向量, 第一层 SNIF-GAT 使用 user2vec 生成的初始用户特征向量 h_u^0 。 $\widetilde{\text{ATT}}$ 是归一化的注意力邻接矩阵, $\mathbf{w}$ 、 $\mathbf{b}$ 是第 $l+1$ 层网络需要训练的权重向量和偏置向量, σ 是网络使用的激活函数。

4 实验与分析

4.1 数据集与实验设置

本文使用的社交网络讽刺文本数据集 SPIRS, 数据内容取自 Twitter 平台, 是目前已开源的含有社交行为信息的高质量讽刺文本数据集, 由 Shmueli 在 2020 年发布。SPIRS 共包含 15 000 条讽刺文本, 同时随机收集了 15 000 条非讽刺文本。本文方法的用户特征向量需要使用用户历史发帖, 因此通过爬虫抓取了各个用户的历史推特, 累计 200 000 条用户历史推特。本文讽刺检测框架使用 2 层 SNIF-GAT, 学习率 $\text{lr}=1 \times 10^{-4}$ 。实验的硬件环境为 NVIDIA TITAN RTX 显卡, 16 GB 运行内存, 处理器使用 Intel Core i5 10100, 实验验证代码在 PyTorch 框架下完成。为了验证方法的有效性并与基准模型进行公平对比, 实验采用和对比模型相同的评价指标, 包括准确率和 $F1$ 值。接下来, 将本文提出的 SNIF-GAT 模型与如下三类模型进行对比实验。

(1) 基于经典神经网络的讽刺检测模型

这类模型使用经典神经网络结构, 只能挖掘文本表层语义信息, 包括经典卷积神经网络 CNN^[8]与使用了注意力机制的 Att-LSTM^[9]。

(2) 基于文本深层语义的讽刺检测模型

这类模型使用更为复杂的神经网络结构, 考虑到讽刺文本的特性, 可以在不同程度上挖掘文

本深层语义。模型包括基于预训练模型的 BERT^[10]、RoBERTa^[11], 其中 RoBERTa 在 BERT 的基础上进一步提升了模型鲁棒性; 基于文本情感冲突的 MIARN^[12]; 基于图神经网络结构的 ADGCN^[16]、ARGAT^[17], 其中 ARGAT 在 ADGCN 的基础上引入了注意力机制。

(3) 基于附加非文本信息的讽刺检测模型

这类模型不仅使用文本语义, 也引入用户历史发帖、社交行为等不同形式的附加信息辅助讽刺检测。模型包括基于用户历史发帖的 CASCADE^[13]、基于用户社交行为的 GAT^[18]。

4.2 对比实验

在 SPIRS 数据集上, 模型对比实验结果如表 2 所示。通过实验验证, SNIF-GAT 在 SPIRS 数据集的 $F1$ 值为 85.62%, 准确率 (Acc) 为 85.33%, 模型性能明显优于已有的文本讽刺检测模型。SPIRS 数据集包含大量社交行为信息, SNIF-GAT 可以充分使用这些社交行为信息作为附加信息, 而这些信息无法被基于文本语义的模型利用, 这是 SNIF-GAT 取得性能优势的关键。

表 2 模型对比实验结果
Table 2 Experimental results of each model

模型	Acc	$F1$
CNN	61.45%	61.32%
Att-LSTM	63.45%	63.77%
BERT	74.45%	74.62%
RoBERTa	75.15%	75.21%
MIARN	72.45%	72.57%
ADGCN	77.35%	77.85%
ARGAT	76.44%	76.32%
GAT	83.15%	83.51%
CASCADE	73.61%	73.42%
SNIF-GAT	85.33%	85.62%

基于附加非文本信息的 GAT 在 SPIRS 上的 $F1$ 值和准确率分别为 83.51%、83.15%, 仅次于 SNIF-GAT。由于 GAT 也可以利用用户社交行为信息, 因此模型性能相较于基于文本语义的模型也有明显领先, 证明了将社交行为作为附加信息的有效性。由于 GAT 直接使用 BERT 生成文本特征向量, 缺乏深层语义的挖掘能力, 且 GAT 在同

质图中使用的注意力机制对不同类型节点信息的区分能力有限,因此 $F1$ 值比 SNIF-GAT 低 2.11%。CASCADE 将用户的历史发帖作为附加信息,但模型无法通过社交网络更新用户特征向量,且模型无法挖掘文本深层语义,因此模型的整体表现大幅低于 SNIF-GAT。

讽刺是一种复杂隐式情感,基于经典神经网络的检测模型无法挖掘深层语义,这类模型在 SPIRS 上表现较差,而基于文本深层语义的检测模型的性能均有明显提升。如表 2 所示,结构复杂的 ADGCN 的 $F1$ 值为 77.85%,在基于文本语义的检测模型中取得了最好表现。但是模型无法检测需要结合用户社交行为信息进行判断的文本,因此在 SPIRS 上的性能低于 SNIF-GAT。

4.3 消融实验

SNIF-GAT 模型融合了文本语义与社交行为信息,为了验证不同模块对模型性能的影响,本文围绕文本特征向量、用户特征向量设计了 2 组消融实验。

在文本特征向量计算中,SNIF-GAT 可以根据任务需求调整情感关键词节点的数量 i ,不同情感关键词数量 i 下的消融实验结果如表 3 所示,实验表明,增加情感关键词节点的数量可以提升模型性能。 $i=0$ 时模型不使用情感关键词节点,此时 $F1$ 值相比基准模型下降了 4.93%,证明了情感子图和情感关键词的有效性。 $i=30$,由于选取的情感关键词数量较少,深层语义挖掘能力偏弱,模型的性能和不使用深层语义的 GAT 接近。 $i=50\sim 100$,情感关键词可以覆盖常见的情感冲突词对, i 的增加可以小幅提升模型性能。

在用户特征向量计算中,基准 SNIF-GAT 模型使用基于边和节点的双重注意力机制,模型在消融实验中可以单独使用基于节点或边的单重注意力机制,也可以不使用注意力机制。SNIF-GAT 使用了节点和边双重注意力机制,基准模型在不同注意力机制下的实验结果如表 4 所示,表中 no attention 为不使用注意力机制的 SNIF-GAT 模型,normal 为完整使用基于边和节点双重注意力机制的 SNIF-GAT 模型,而 node 和 edge 代表分别只使用节点注意力和边注意力的 SNIF-GAT 模型。实验结果表明,节点注意力和边注意力均对模型性能产生重要影响。

表 3 不同情感关键词数量下的实验结果
Table 3 Experimental result of different numbers of affective keywords

关键词数量	Acc	$F1$ 值
$i=0$	80.87%	80.69%
$i=30$	83.31%	83.66%
$i=50$	85.33%	85.62%
$i=80$	85.45%	85.51%
$i=100$	85.62%	85.77%
$i=150$	85.65%	85.76%
$i=200$	85.68%	85.78%

表 4 不同注意力机制下的实验结果
Table 4 Experimental results of different attention mechanism

注意力机制	Acc	$F1$ 值
no attention	78.21%	78.46%
edge	83.73%	83.95%
node	83.45%	83.62%
normal	85.33%	85.62%

在不使用注意力机制的情况下,模型的用户特征向量为 user2vec 生成的初始特征向量,无法在社交网络中进行表征更新。此时模型 $F1$ 值相比于基准模型下降了 7.16%,证明了双重注意力机制的有效性。在单独使用节点注意力和边注意力的情况下,二者相比基准模型的 $F1$ 值均下降 2% 左右,证明了节点注意力和边注意力的有效性。

4.4 模型超参数分析

SNIF-GAT 的层数和特征向量的维度 d 是两个较为重要的超参数,下面通过实验进行超参数性能分析。

为了验证 SNIF-GAT 层数对模型性能的影响,本文分别对 1、2、3、4 层 SNIF-GAT 进行性能测试,不同 SNIF-GAT 层数下的实验结果如表 5 所示。实验表明,使用 2 层 SNIF-GAT 时可以取得最佳表现,继续增加网络层数会导致用户特征向量趋于一致,节点的区分度下降,进而导致模型性能下降。

表 5 不同 SNIF-GAT 层数下的实验结果
Table 5 Experimental results of different SNIF-GAT depth

SNIF-GAT 层数	Acc	$F1$ 值
1	82.21%	82.36%
2	85.33%	85.62%
3	83.15%	83.41%
4	79.19%	79.47%

不同特征向量维度 d 下的实验结果如图 6 所示, 实验表明, $d=300$ 时模型获得最佳表现。当 d 较小时, 特征向量对于复杂信息的表征能力不足, 增加向量维度可以提升模型性能。当 $d>300$ 时, 继续增加特征向量维度可能会造成模型过拟合, 进而降低模型的性能。实验中 d 取 400 和 500 时, 模型的 $F1$ 值和准确率均出现了降低。

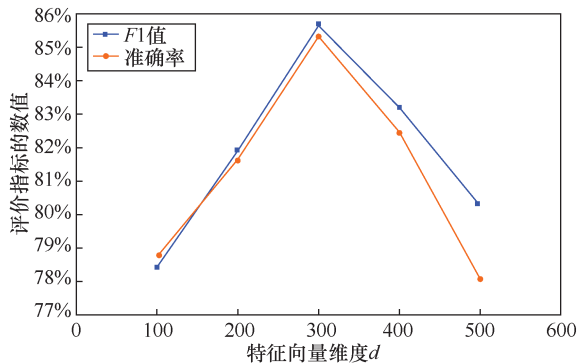


图 6 不同特征向量维度 d 下的实验结果

Figure 6 Experimental result under different feature vectors dimension

4.5 样例分析

为进一步分析模型在真实讽刺文本中的表征能力, 本文分别在显式讽刺和隐式讽刺的样例文本上进行了测试, 文本特征向量和用户特征向量的可视化分析如表 6 所示。

表 6 文本特征向量和用户特征向量的可视化分析
Table 6 Visualization of the text and user embedding

类型	模型	文本特征向量	用户特征向量
显示讽刺	SNIF-GAT	I love to be ignored	讽刺倾向
	ADGCN	I love to be ignored	无
	GAT	I love to be ignored	讽刺倾向
隐式讽刺	SNIF-GAT	Huh, I really love this big dinner	讽刺倾向
	ADGCN	Huh, I really love this big dinner	无
	GAT	Huh, I really love this big dinner	讽刺倾向

表 6 中文本特征的背景颜色深浅表示注意力权重大小, 用户特征中背景颜色表示该用户在社交文本中使用讽刺表达的倾向性, 大小通过用户特征向量在分类器中分类为讽刺的概率表示。在显示讽刺文本中, SNIF-GAT 和 ADGCN 都准确捕捉到 “love” 和 “ignore” 的情感冲突, GAT 使用 BERT 生成文本特征向量, 无法进行深层语义分析。在隐式讽刺文本中, 例句没有明显情感冲突, SNIF-GAT 和 ADGCN 都捕捉到了正面情感词, 单纯依靠文本特征分类将得到错误结果。但

由于 SNIF-GAT 使用情感关键词, 文本特征向量中引入的噪声信息较少, 通过聚合关联节点的信息, SNIF-GAT 生成的用户特征向量表示该用户有较高的讽刺倾向, 在与文本特征向量融合后, 分类器最终准确识别该文本为讽刺类型。

5 结束语

针对社交网络环境中讽刺文本检测问题, 本文提出了一种基于文本语义和社交行为信息融合的讽刺检测方法。本文针对社交网络文本的特点构造了包含用户、文本、情感关键词的社交异质信息网络 (UTE-HIN), 并通过 SNIF-GAT 模型对 UTE-HIN 进行表征计算, 得到包含文本语义和社交行为信息的融合特征向量。融合特征向量用作分类器的输入, 可以实现对文本是否为讽刺类型的检测。通过在公开数据集上的对比实验表明, 本文方法相较现有方法在分类效果上有显著提升。未来研究可以综合利用社交网络中的多媒体信息以及更为复杂的用户交互关系, 通过更丰富的社交行为信息进一步提升讽刺检测的水平。

参考文献:

- [1] VERMA P, SHUKLA N, SHUKLA A P. Techniques of sarcasm detection: a review[C]//2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE). 2021: 968-972.
- [2] JIA X, DENG Z, MIN F, et al. Three-way decisions based feature fusion for chinese irony detection[J]. International Journal of Approximate Reasoning, 2019, 113: 324-335.
- [3] CAI Y, CAI H, WAN X. Multi-modal sarcasm detection in twitter with hierarchical fusion model[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 2506-2515.
- [4] SHRIVASTAVA M, KUMAR S. A pragmatic and intelligent model for sarcasm detection in social media text[J]. Technology in Society, 2021, 64: 101489.
- [5] RILOFF E, QADIR A, SURVE P, et al. Sarcasm as contrast between a positive sentiment and negative situation[C]//Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. 2013: 704-714.
- [6] BOUAZIZI M, OHTSUKI T O. A pattern-based approach for sarcasm detection on twitter[J]. IEEE Access, 2016, 4: 5477-5488.
- [7] AMIR S, WALLACE B C, LYU H, et al. Modelling context with user embeddings for sarcasm detection in social media[C]//Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning. 2016: 167-177.
- [8] DAS D, CLARK A J. Sarcasm detection on flickr using a

- cnn[C]//Proceedings of the 2018 International Conference on Computing and Big Data. 2018: 56-61.
- [9] KUMAR A, NARAPAREDDY V T, SRIKANTH V A, et al. Sarcasm detection using multi-head attention based bidirectional LSTM[J]. IEEE Access, 2020, 8: 6388-6397.
- [10] BARUAH A, DAS K, BARBHUIYA F, et al. Context-aware sarcasm detection using bert[C]//Proceedings of the Second Workshop on Figurative Language Processing. 2020: 83-87.
- [11] GREGORY H, LI S, MOHAMMADI P, et al. A transformer approach to contextual sarcasm detection in Twitter[C]//Proceedings of the 2nd Workshop on Figurative Language Processing. 2020: 270-275.
- [12] TAY Y, TUAN L A, HUI S C, et al. Reasoning with sarcasm by reading in-between[J]. arXiv Preprint arXiv:1805.02856, 2018.
- [13] HAZARIKA D, PORIA S, GORANTLA S, et al. CASCADE: contextual sarcasm detection in online discussion forums[C]//Proceedings of the 27th International Conference on Computational Linguistics. 2018: 1837-1848.
- [14] KIPF T N, WELING M. Semi-supervised classification with graph convolutional networks[J]. arXiv Preprint arXiv:1609.02907, 2016.
- [15] 陈皓, 易平. 基于图神经网络的代码漏洞检测方法[J]. 网络与信息安全学报, 2021, 7(3): 37-45.
- CHEN H, YI P. Code vulnerability detection method based on graph neural network[J]. Chinese Journal of Network and Information Security, 2021, 7(3): 37-45.
- [16] LOU C, LIANG B, GUI L, et al. Affective dependency graph for sarcasm detection[C]//Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2021: 1844-1849.
- [17] LI G, LIN F, CHEN W, et al. Affection enhanced relational graph attention network for sarcasm detection[J]. Applied Sciences, 2022, 12(7): 3639.
- [18] PLEPI J, FLEK L. Perceived and intended sarcasm detection with graph attention networks[C]//Findings of the Association for Computational Linguistics. 2021: 4746-4753.
- [19] SHMUELI B, KU L W, RAY S. Reactive supervision: a new method for collecting sarcasm data[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020: 2553-2559.
- [20] CAMBRIA E, LI Y, XING F Z, et al. SenticNet 6: ensemble application of symbolic and sub-symbolic AI for sentiment analysis[C]//Proceedings of the 29th ACM International Conference on Information & Knowledge Management. 2020: 105-114.
- [21] HALLAC I R, MAKINIST S, AY B, et al. user2vec: social media user representation based on distributed document embeddings[C]//2019 International Artificial Intelligence and Data Processing Symposium. 2019: 1-5.
- [22] GOEL A, GAUTAM J, KUMAR S. Real time sentiment analysis of tweets using Naive Bayes[C]//2016 2nd International Conference on Next Generation Computing Technologies (NGCT). 2016: 257-261.
- [23] RAMCHOUN H, GHANOU Y, ETTAOUIL M, et al. Multilayer perceptron: architecture optimization and training[R]. 2016.
- [24] ZHANG C, LI Q C, SONG D W. 2019. Aspect-based sentiment classification with aspect-specific graph convolutional networks[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019: 4568-4578.
- [25] OPREA S, MAGDY W. Exploring author context for detecting intended vs perceived sarcasm[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 2854-2859.
- [26] KOLCHINSKI Y A, POTTS C. Representing social media users for sarcasm detection[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018: 1115-1121.
- [27] YANG T, HU L, SHI C, et al. HGAT: heterogeneous graph attention networks for semi-supervised short text classification[J]. ACM Transactions on Information Systems (TOIS), 2021, 39(3): 1-29.

[作者简介]



付朝阳 (1995—), 男, 江苏徐州人, 上海交通大学硕士生, 主要研究方向为社交网络大数据分析。



陈致凯 (1998—), 男, 上海人, 上海交通大学硕士生, 主要研究方向为图表示学习、数据挖掘。



潘理 (1974—), 男, 江苏苏州人, 博士, 上海交通大学教授、博士生导师, 主要研究方向为网络大数据分析、云计算与大数据安全、网络安全管理。