山西大学学报(自然科学版)49(4):581-591,2026
Journal of Shanxi University(Nat. Sci. Ed.)

DOI:10.13451/j.sxu.ns.2025121

基于常识知识和情绪知识语义不一致的讽刺检测方法

江浩,万中英,曾雪强*,王明文
(江西师范大学 人工智能学院,江西 南昌 330022)

摘 要:讽刺是一种文本字面含义与真实情感意图之间存在不一致的修辞手法,准确建模语义不一致对于实现精准的讽刺检测至关重要。语义不一致不仅存在于文本本身的表述中,更与常识知识和情绪知识密切相关。然而,现有研究一般仅局限于单独探讨文本与常识知识或文本与情绪知识之间的语义不一致,缺乏同时考虑文本与两种外部知识的语义不一致。针对上述问题,该文提出了一种基于常识知识和情绪知识语义不一致的讽刺检测方法(Semantic Incongruity Incorporating Commonsense and Emotional Knowledge for Sarcasm Detection, SIICE)。SIICE模型采用注意力机制对文本语义、常识知识和情绪知识相互之间存在的语义不一致进行建模,并融合三种不一致信息进行讽刺检测。SIICE模型包含四个模块:文本语义不一致感知模块、常识知识语义不一致感知模块、情绪知识语义不一致感知模块和融合预测模块。其中,前三个模块分别利用注意力机制获取文本内部、文本与常识转换器(Commonsense Transformers,COMET)生成的常识知识之间,以及文本与效价、唤醒度与支配度(Valence-Arousal-Dominance,VAD)情绪模型生成的情绪知识之间的语义不一致表示;融合预测模块将上述三种表示拼接后进行讽刺检测。为了验证SIICE模型的有效性,该文在Twitter(Ptáček)、IAC-V1和IAC-V2三个基准数据集上进行实验评估,结果表明SIICE模型在讽刺检测任务中的性能优于多个基线模型,*F*1值分别达到0.870 3、0.704 9和0.783 5。消融实验表明,同时建模文本与常识知识及情绪知识之间的语义不一致可以有效提升讽刺检测模型的性能。

关键词:讽刺识别;语义冲突;外部知识;情感分布
中图分类号:TP39 **文献标志码**:A **文章编号**:0253-2395(2026)04-0581-11

Semantic Incongruity Incorporating Commonsense and Emotional Knowledge for Sarcasm Detection

Jiang Hao, Wan Zhongying, Zeng Xueqiang*, Wang Mingwen
(*School of Artificial Intelligence, Jiangxi Normal University, Nanchang* 330022, *China*)

Abstract: Sarcasm involves semantic incongruity between literal meanings and intended emotions. Accurate modeling of such incongruity is critical for sarcasm detection. Semantic incongruity occurs not only within textual content but also between the text and external commonsense and emotional knowledge. Previous studies typically considered semantic incongruity either between text and commonsense or text and emotional knowledge separately, neglecting their joint influence. This study proposed a sarcasm detection model named Semantic Incongruity Incorporating Commonsense and Emotional Knowledge (SIICE). The SIICE model employed an attention mechanism to model semantic incongruities among textual semantics, commonsense knowledge, and emotional knowledge. SIICE consisted of four modules: a textual semantic incongruity module, a commonsense semantic incongruity module, an emotional semantic incongruity module, and a fusion prediction module. The first three modules utilized attention mechanisms to capture semantic incongruities within the text, between text and commonsense knowledge generated by COMET (Commonsense

收稿日期:2025-10-23;**修回日期**:2025-11-24
基金项目:国家自然科学基金(62266021;62266023)
作者简介:江浩(2000—),男,江西贵溪人,硕士研究生,研究方向为自然语言处理。E-mail:jianghao0539@163.com
* **通信作者**:曾雪强(Zeng Xueqiang),E-mail:xqzeng@jxnu.edu.cn
引文格式:江浩,万中英,曾雪强,等.基于常识知识和情绪知识语义不一致的讽刺检测方法[J].山西大学学报(自然科学版),2026,49(4):581-591. DOI:10.13451/j.sxu.ns.2025121.

Transformers), and between text and emotional knowledge represented via the VAD (Valence-Arousal-Dominance) emotional model, respectively. The fusion module integrated these three representations for sarcasm detection. Experiments were conducted on three benchmark datasets: Twitter(Ptáček), IAC-V1, and IAC-V2. SIICE achieved superior performance compared to baseline models, obtaining *F*1-scores of 0.870 3, 0.704 9, and 0.783 5, respectively. Ablation experiments indicated that simultaneous modeling of semantic incongruities involving text, commonsense, and emotional knowledge effectively enhances sarcasm detection performance.

Key words: irony detection; semantic conflict; external knowledge; emotion distribution

0 引言

讽刺是一种广泛存在于社交媒体的修辞手法,通过隐含的嘲讽或轻蔑态度表达与字面含义相反的真实意图[1]。讽刺检测的关键在于识别文本中存在的语义不一致。简单讽刺通常表现为浅层语义冲突,而复杂讽刺则涉及文本之外的常识知识与情绪知识等背景信息,更加隐晦且更难识别[2]。其中,常识知识是指人类在长期社会实践中普遍形成、并用以理解日常情境的背景信息[3];情绪知识是人们对情绪及其表达的共同认知,包括对情绪表现、反应及社交变化的普遍理解[4]。二者在揭示复杂讽刺表达中的隐含语义冲突方面均具有重要作用。如图1所示,句子"I love to see a doctor every day"字面表达积极情绪("love"),仅凭文本难以察觉其讽刺意图。引入常识知识后可以发现"see a doctor"通常与"痛苦(painful)""焦虑(anxious)""紧张(nervous)"等负面体验相关联,这与表面积极情绪构成了语义冲突;从情绪角度分析,个体对于频繁、持续的医疗活动一般不会产生真实的积极情感,进一步揭示了文本表达与真实情感意图之间的不一致。因此,在讽刺检测任务中,同时建模文本语义、常识知识和情绪知识相互之间的不一致,有助于更准确地识别潜在的语义冲突。

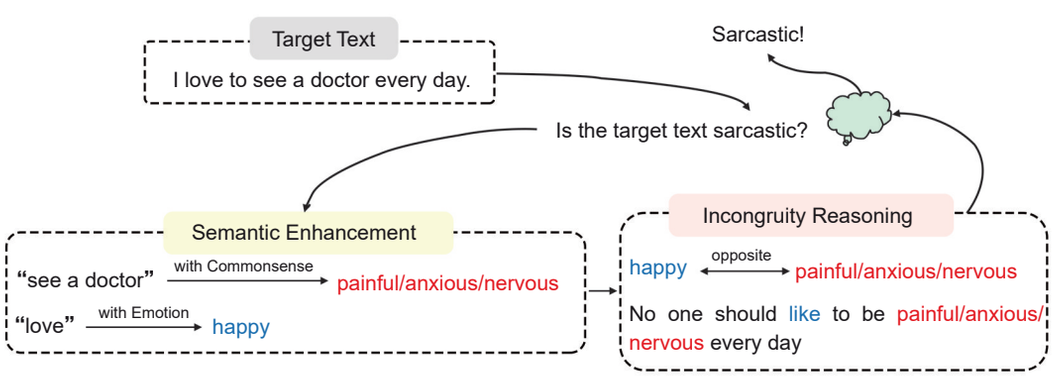


图1 讽刺检测示例

Fig. 1 Example of sarcasm detection

早期讽刺检测研究主要关注文本内部语义不一致特征,通过挖掘文本自身的语义冲突识别讽刺。其中,一类方法侧重于词汇层面的语义建模,如Xiong等[5]通过注意力机制建模词汇间语义关系;另一类研究则关注片段级语义差异,如Pan等[6]采用自注意力机制赋予句内片段不同权重,突出片段语义不一致。然而,讽刺表达通常还受到外部常识知识和内在心理因素的共同影响,仅依靠文本内部语义不一致难以全面准确地捕捉其讽刺意图。

为克服上述局限,近期研究逐渐引入常识或情绪知识,以深入理解文本中的隐含语义及情感倾向。一方面,一些研究融合常识知识强化对文本隐含含义的认知,如Li等[7]借助常识转换器(Commonsense Transformers, COMET)模型生成常识知识,并通过门控与残差机制实现文本和常识知识的融合。此外,Chen等[8]受图模型研究的启发,构建文本与常识的有向异构图,利用多头注意力提取语义信息,但忽略了句法依存关系的重要性;为弥补这一不足,Yu等[9]以常识增强的情感图和依存图为基础,利用图卷积网络和注意力机制融合图特征与句

子表示。

另一方面,情绪知识也逐渐成为讽刺检测研究的关注点。一些研究利用情感特征揭示深层次语义矛盾,例如,Babanejad等[10]利用情感与上下文特征来扩展基于Transformer的双向编码器表征(Bidirectional Encoder Representations from Transformers,BERT)结构用于讽刺检测;此外,Lou等[11]通过图卷积神经网络构建情感图和句法依存图,捕获句子内的长程情感不一致;为了减少人工构建情感图带来的误差,Wang等[12]通过迭代的图学习过程对情感图和依存图进行扩充,以获得优化的语义不一致图结构。

尽管上述研究分别探讨了文本与常识知识或情绪知识之间的语义不一致,但缺乏同时考虑文本与两种外部知识之间的语义不一致,限制了模型在更复杂、隐蔽的真实场景中准确识别讽刺意图的能力。针对上述问题,本文提出了一种基于常识知识和情绪知识语义不一致的讽刺检测方法(Semantic Incongruity Incorporating Commonsense and Emotional Knowledge for Sarcasm Detection, SIICE)。SIICE模型采用注意力机制对文本语义、常识知识和情绪知识相互之间的不一致进行建模,并融合三种不一致信息以及文本自身语义表示进行讽刺检测。SIICE模型包括四个模块:(1)文本语义不一致感知模块,利用自注意力机制捕捉文本内部的语义不一致表示;(2)常识知识语义不一致感知模块,利用交互注意力机制建模文本与COMET生成的常识知识之间的语义不一致表示;(3)情绪知识语义不一致感知模块,采用交互注意力机制捕获文本与效价、唤醒度与支配度(Valence-Arousal-Dominance,VAD)情绪模型生成的情绪知识之间的语义不一致表示;(4)融合预测模块,融合上述三种跨域语义不一致表示并与文本语义表示进行拼接,最后通过全连接网络检测讽刺。

1 相关工作

1.1 基于文本的讽刺检测方法

传统的讽刺检测方法主要依赖于文本语义信息,可进一步划分为无上下文与基于上下文两类方法。无上下文方法仅分析目标语句自身语义,而上下文相关方法则考虑目标语句与语境的语义关系[13]。如Tay等[14]提出一种基于注意力机制的方法捕捉语义冲突,部分研究则通过迁移学习引入情绪识别领域的知识,以提升模型的泛化能力。

然而,讽刺的识别通常依赖于更广泛的语境,包括说话者状态、文本所处的语境位置以及社交媒体互动信息等。Poria等[15]利用预训练模型提取用户个性与情绪特征以增强讽刺识别效果;Hazarika等[16]提出上下文讽刺检测器(Contextual Sarcasm Detector,CASCADE)模型,通过联合建模文本与多维上下文(用户属性、主题信息)进一步提升讽刺检测性能。然而,随着社交媒体表达方式的多样化,仅凭文本信息难以全面满足复杂讽刺情景下对精准语义与语境建模的需求。

1.2 基于常识知识的讽刺检测方法

近年来,讽刺检测研究逐渐从纯文本语义建模转向显式引入外部常识知识,以揭示更隐蔽的语境与语义矛盾。早期研究提出了词汇层面的常识知识库ConceptNet[17]和事件层面的社会常识知识图谱(An Atlasof Machine Commonsense,ATOMIC)[18]。其中,ATOMIC定义了九种常识关系(见表1),并可以形式化为三元组(s,r,o),其中s是主语短语,r是关系短语,o是宾语短语。随后,Bosselut等[19]提出了一种基于生成式预训练Transformer(Generative Pre-trained Transformer, GPT)的常识生成模型COMET,以ATOMIC为基础训练的COMET能够根据给定输入生成九类常识知识。

近期研究进一步探索基于图结构的常识融合机制。如Yu等[9]将COMET生成的常识知识融入情感图与句法依存图中,利用图卷积网络与注意力机制整合图结构特征和句子语义表示;为了增强模型常识推理能力与知识完备性,Qiu等[20]利用检索增强的GPT-4o补全常识,结合优化依存图和情感不一致推理,通过层次化图注意力提取关键语义特征。然而,上述研究仅关注文本与常识知识之间的语义不一致,忽略了文本与情绪知识之间的语义冲突,限制了模型对更隐晦的讽刺现象的感知和泛化能力。

表1 ATOMIC常识关系、含义和解释

Table 1 ATOMIC commonsense relationships, meanings and explanations

常识关系	含义	解释
xIntent	说话者的意图	X做出该事件的意图
xNeed	说话者的需求	事件发生前X的需求
xAttr	说话者的属性	对X的属性描述
xEffect	对说话者的影响	事件对X产生的影响
xWant	说话者想要的	事件发生后X产生的意图
xReact	说话者的反应	事件发生后X的反应
oReact	听者的反应	事件发生后他人的反应
oWant	听者想要的	事件发生后他人产生的意图
oEffect	听者的影响	事件对他人产生的影响

1.3 基于情绪知识的讽刺检测方法

在讽刺检测任务中,情绪知识是指来自情绪词典或情绪常识图谱的、可检索的先验情感信息,用来度量情绪取向的不一致,帮助模型识别讽刺话语[12]。如Russell等[21]提出的VAD情绪模型,将情绪表征为空间连续分布的三个维度,即效价(Valence)、唤醒(Arousal)与支配(Dominance),实现可量化表达。

近年来,一类方法借助多任务学习架构,利用情感分析任务辅助讽刺检测。例如,Savini等[22]基于双向长短期记忆网络(Bidirectional-Long Short-term Memory,BiLSTM)构建多任务学习框架,共享嵌入层并通过特定的多层感知机协同优化讽刺检测主任务与情感分类辅助任务。此外,Chauhan等[23]将情感分析作为次要任务,提出了跨模态与模态内不一致信息融合的注意力机制。另一类研究则直接融入情感特征来强化文本表征,如Babanejad等[10]在BERT架构中融入情感与上下文特征来进行讽刺检测。Agrawal等[24]将讽刺检测建模为序列分类问题,利用文本内在的情感状态转换以提高识别性能。然而,上述方法虽然关注了情绪知识与文本之间的语义不一致,却忽略了常识知识对讽刺识别的重要性。同时考虑文本与两种外部知识的语义不一致,才能更深入地捕捉语义不协调的本质特征,实现对讽刺现象的精准识别。

2 基于常识知识和情绪知识语义不一致的讽刺检测方法

本文针对文本与常识知识、文本与情绪知识之间存在的语义不一致,提出了一种基于常识知识和情绪知识语义不一致的讽刺检测方法(SIICE)。SIICE模型采用注意力机制对文本语义、常识知识和情绪知识相互之间的不一致进行建模,融合三种不一致信息并与文本语义表示拼接进行讽刺检测。SIICE模型由四个模块组成:文本语义不一致感知模块、常识知识语义不一致感知模块、情绪知识语义不一致感知模块和融合预测模块。SIICE模型的具体架构如图2所示。

2.1 文本语义不一致感知模块

2.1.1 文本表示

本文采用预训练RoBERTa模型提取文本表示。给定一个由n个词组成的文本序列$w=\{w_1,w_2,\cdots,w_n\}$,首先在序列起始位置添加特殊标记[CLS],构造模型输入序列。随后,将该序列输入RoBERTa模型,得到文本嵌入表示:

$$\boldsymbol{h}_{\mathrm{CLS}},\boldsymbol{H}_o=\mathrm{RoBERTa}\left(\left[[\mathrm{CLS}];w\right]\right),\quad(1)$$

其中$\boldsymbol{h}_{\mathrm{CLS}}\in R^{1\times l_c}$是RoBERTa模型最后一层中[CLS]标记对应的向量表示。$\boldsymbol{H}_o=\{\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_n\}$是除[CLS]标记之外的其他位置的表示,$\boldsymbol{H}_o\in R^{n\times l_c}$,$l_c$为隐藏层维度。

2.1.2 获取文本语义不一致表示

为了捕捉文本内部存在的语义不一致,本文引入了注意力机制进行建模。已有研究表明,注意力机制能够有效关注并识别文本中的语义不一致现象,并进行讽刺检测[1,5-6,14]。

首先,将向量$\boldsymbol{h}_{\mathrm{CLS}}$作为注意力机制的查询(Query),文本嵌入表示$\boldsymbol{H}_o$作为键矩阵(Key)和值矩阵(Value)。然后,将$\boldsymbol{h}_{\mathrm{CLS}}$与$\boldsymbol{H}_o$进行内积运算,并使用$f_{\mathrm{soft\,max}}$函数进行归一化得到文本中每个词嵌入表示对应的注意力得分,最后将其与$\boldsymbol{H}_o$中每个词嵌入表示进行加权求和,得到文本语义不一致表示$\boldsymbol{i}_t\in R^{1\times l_c}$:

$$\boldsymbol{i}_t=\sum_{j=1}^{n}f_{\mathrm{softmax}}\left(\frac{\boldsymbol{h}_{\mathrm{CLS}}\boldsymbol{h}_j{}^T}{\sqrt{l_c}}\right)\boldsymbol{h}_j,\quad(2)$$

其中$\boldsymbol{h}_j$是$\boldsymbol{H}_o$的第j行;l_c表示向量维度,$\sqrt{l_c}$主要为防止内积值随着向量的维度增大而增加,进而使梯度趋于稳定。

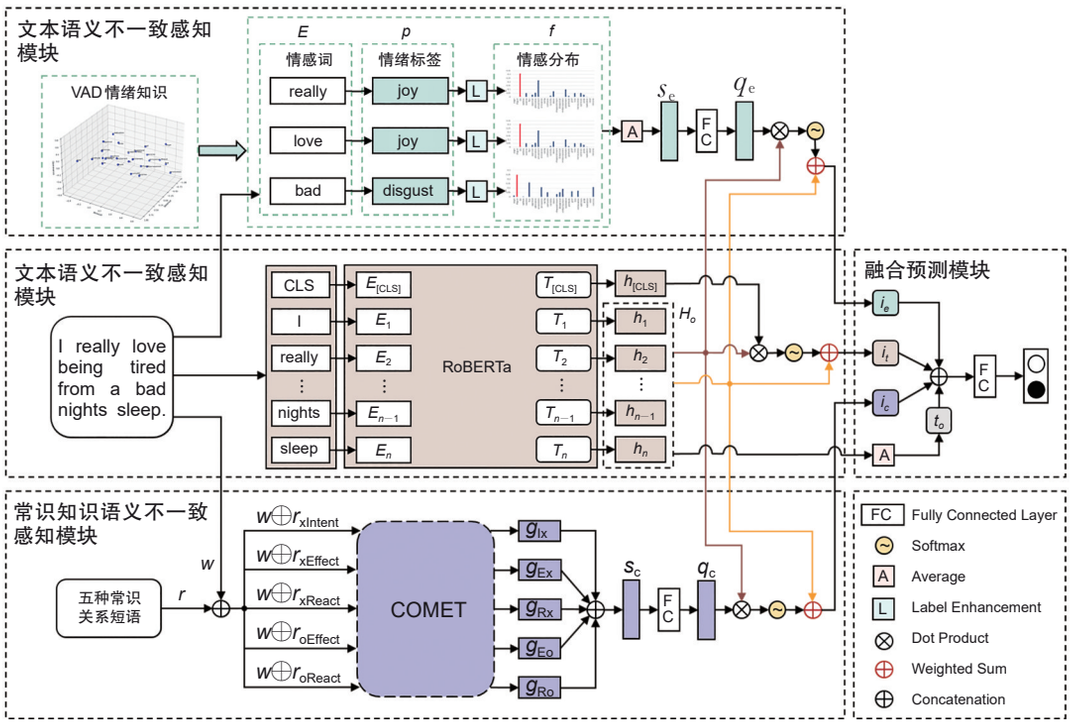


图2 SIICE模型框架图

Fig. 2 Structure of SIICE Model

2.2 常识知识语义不一致感知模块

2.2.1 生成多角度常识知识

本文采用在ATOMIC知识图谱上预训练的COMET模型获取句子中的常识信息[19]。COMET利用ATOMIC中的三元组(s,r,o)进行训练,根据拼接的主语短语s和关系短语r生成相应的宾语短语o。本文从表1中选取五种常识关系:说话者的意图(xIntent)、对说话者的影响(xEffect)、说话者的反应(xReact)、对听者的影响(oEffect)和听者的反应(oReact),用于生成对应的常识知识。相关研究表明,这五种常识关系涵盖了说话者内在的意图、事件的影响以及双方的情绪反应,能够为模型提供较为全面且平衡的视角[7,25]。

首先,将输入文本w作为主语短语分别与五种常识关系短语r拼接,构造模型的输入序列;COMET根据不同常识关系类型的输入序列生成相应的常识知识向量表示$\boldsymbol{g}_{Ix}$、$\boldsymbol{g}_{Ex}$、$\boldsymbol{g}_{Rx}$、$\boldsymbol{g}_{Eo}$和$\boldsymbol{g}_{Ro}$:

$$\begin{aligned}\boldsymbol{g}_{Ix}&=\mathrm{COMET}\left(w\oplus r_{\mathrm{xIntent}}\right),\\ \boldsymbol{g}_{Ex}&=\mathrm{COMET}\left(w\oplus r_{\mathrm{xEffect}}\right),\\ \boldsymbol{g}_{Rx}&=\mathrm{COMET}\left(w\oplus r_{\mathrm{xReact}}\right),\\ \boldsymbol{g}_{Eo}&=\mathrm{COMET}\left(w\oplus r_{\mathrm{oEffect}}\right),\\ \boldsymbol{g}_{Ro}&=\mathrm{COMET}\left(w\oplus r_{\mathrm{oReact}}\right)。\end{aligned}\tag{3}$$

其中$\oplus$是拼接操作;r_{xIntent}、r_{xEffect}、r_{xReact}、r_{oEffect}和r_{oReact}分别代表常识关系xIntent、xEffect、xReact、oEffect和oReact对应的关系短语。

最后,将这五个向量拼接成文本的综合常识知识向量表示$\boldsymbol{s}_c\in R^{1\times(5l_c)}$:

$$\boldsymbol{s}_c=\boldsymbol{g}_{Ix}\oplus\boldsymbol{g}_{Ex}\oplus\boldsymbol{g}_{Rx}\oplus\boldsymbol{g}_{Eo}\oplus\boldsymbol{g}_{Ro}。\tag{4}$$

2.2.2 获取常识知识语义不一致表示

本文利用交互注意力机制获取常识知识语义不一致表示,该机制将常识知识向量作为查询,与文本嵌入表示矩阵进行交叉注意力计算,从而捕捉文本与常识知识之间的语义不一致。

首先,将常识知识向量$\boldsymbol{s}_c$输入全连接网络(f_{FC}),降维得到与隐藏层维度一致的向量$\boldsymbol{q}_c=f_{\mathrm{FC}}(\boldsymbol{s}_c)\in R^{1\times l_c}$。$\boldsymbol{q}_c$作为注意力机制的查询(Query),文本嵌入表示$\boldsymbol{H}_o$作为键矩阵(Key)和值矩阵(Value)。然后,计算$\boldsymbol{q}_c$与$\boldsymbol{H}_o$之间的内积,并利用$f_{\mathrm{soft\,max}}$函数进行归一化得到每个词嵌入表示对应的注意力得分,将其与$\boldsymbol{H}_o$中每个词嵌入表示$\boldsymbol{h}_j$进行加权求和,得到常识知识语义不一致表示$\boldsymbol{i}_c\in R^{1\times l_c}$:

$$\boldsymbol{i}_{\mathrm{c}}=\sum_{j=1}^{n} f_{\mathrm{softmax}}\left(\frac{\boldsymbol{q}_{\mathrm{c}}\boldsymbol{h}_j^{T}}{\sqrt{l_{\mathrm{c}}}}\right)\boldsymbol{h}_j。\tag{5}$$

2.3 情绪知识语义不一致感知模块

2.3.1 生成情绪标签的情感分布

本文基于VAD情绪模型中的情绪距离,为句子中情感词E的情绪标签p生成相应的情感分布$\boldsymbol{f}$,再通过均值化操作得到句子的综合情感分布表示向量$\boldsymbol{s}_e$,具体计算公式如下[26-27]:

$$\boldsymbol{f}_{\alpha}^{e}=\frac{1}{Z}\exp\left(-\frac{(\left\|\boldsymbol{\mu}_e-\boldsymbol{\mu}_{\alpha}\right\|_2+b)^2}{2\tau^2}\right),\tag{6}$$

$$Z=\sum_{e}\exp\left(-\frac{(\left\|\boldsymbol{\mu}_e-\boldsymbol{\mu}_{\alpha}\right\|_2+b)^2}{2\tau^2}\right),\tag{7}$$

$$\boldsymbol{s}_{\mathrm{e}}=\frac{1}{m}\cdot\frac{1}{h}\cdot\sum_{k=1}^{m}\sum_{t=1}^{h}\boldsymbol{f}_{p_k^t}。\tag{8}$$

其中$\boldsymbol{\mu}_{\mathrm{e}}=\left[V^{\mathrm{e}},A^{\mathrm{e}},D^{\mathrm{e}}\right]$表示情绪$e$的VAD均值向量;偏置项$b$取值为1;局部加权带宽参数$\tau$用于调节情感分布散布程度;归一化因子$Z$保证$\sum_{\mathrm{e}}\boldsymbol{f}_{\alpha}^{\mathrm{e}}=1$;$\left\|\boldsymbol{\mu}_{\mathrm{e}}-\boldsymbol{\mu}_{\alpha}\right\|_2$表示情绪$e$与$\alpha$在VAD空间中的欧氏距离;$m$与$h$分别记为句子中情感词数量及其关联的情绪标签数量;$\boldsymbol{f}_{p_k^t}$则表示第$k$个情感词的第$t$个情绪标签$p_k^t$对应的情感分布。

2.3.2 获取情绪知识语义不一致表示

为了获取情绪知识语义不一致表示,本文采用交互注意力机制,将情绪知识向量作为查询,与文本嵌入表示矩阵进行交叉注意力计算,从而捕捉文本与情绪知识之间的语义不一致。

首先,将情感分布表示向量$\boldsymbol{s}_{\mathrm{e}}$输入全连接网络($f_{\mathrm{FC}}$),映射为与隐藏层维度一致的向量$\boldsymbol{q}_{\mathrm{e}}=f_{\mathrm{FC}}(\boldsymbol{s}_{\mathrm{e}})\in R^{1\times l_{\mathrm{c}}}$。$\boldsymbol{q}_{\mathrm{e}}$作为注意力机制的查询(Query),文本嵌入表示$\boldsymbol{H}_o$作为键矩阵(Key)和值矩阵(Value)。然后,计算$\boldsymbol{q}_{\mathrm{e}}$与$\boldsymbol{H}_o$之间的内积,并使用$f_{\mathrm{softmax}}$函数进行归一化得到每个词嵌入表示对应的注意力得分,将其与$\boldsymbol{H}_o$中每个词嵌入表示$\boldsymbol{h}_j$进行加权求和,得到情绪知识语义不一致表示$\boldsymbol{i}_{\mathrm{e}}\in R^{1\times l_{\mathrm{c}}}$:

$$\boldsymbol{i}_{\mathrm{e}}=\sum_{j=1}^{n} f_{\mathrm{softmax}}\left(\frac{\boldsymbol{q}_{\mathrm{e}}\boldsymbol{h}_j^{T}}{\sqrt{l_{\mathrm{c}}}}\right)\boldsymbol{h}_j。\tag{9}$$

2.4 融合预测模块

融合预测模块将文本语义不一致表示$\boldsymbol{i}_{\mathrm{t}}$、常识知识语义不一致表示$\boldsymbol{i}_{\mathrm{c}}$、情绪知识语义不一致表示$\boldsymbol{i}_{\mathrm{e}}$进行融合,并与文本嵌入表示$\boldsymbol{H}_o$的均值$\boldsymbol{t}_o=\frac{1}{\boldsymbol{n}}\sum_{j=1}^{n}h_j\in R^{1\times l_{\mathrm{c}}}$进行拼接。随后,通过全连接网络($f_{\mathrm{FC}}$)和$f_{\mathrm{soft\,max}}$函数输出预测结果$\hat{y}$:

$$\hat{y}=f_{\mathrm{softmax}}\left(f_{\mathrm{FC}}\left(\boldsymbol{i}_{\mathrm{t}}\oplus\boldsymbol{i}_{\mathrm{c}}\oplus\boldsymbol{i}_{\mathrm{e}}\oplus\boldsymbol{t}_o\right)\right)。\tag{10}$$

本模型采用二元交叉熵损失函数作为训练目标,用于计算二分类任务的损失$\mathcal{L}$:

$$L=\left(y^j\log\left(\hat{y}^j\right)\right)+\left(1-y^j\right)\log\left(1-\hat{y}^j\right),\tag{11}$$

其中y^j是真实标签值,$\hat{y}^j$是预测概率值。

3 实验及结果分析

3.1 数据集和实验设置

本文采用三个广泛用于讽刺检测研究的基准数据集评估模型性能,包括两个长文本数据集IAC-V1和IAC-V2(Internet Argument Corpus)[28],以及一个短文本数据集Twitter(Ptáček)[29]。

在数据预处理阶段,为了提高文本质量并减少噪声影响,本文对数据进行了以下处理:将所有文本转为小写;移除HTTP链接;删除包含@的用户标识符;去除与sarcasm相关的标签;删除与语义分析无关的标点符号;并对哈希标签进行分割处理。随后,数据集按照8∶1∶1比例划分为训练集、验证集和测试集。数据集统计结果如表2所示。

本文采用3种度量指标来评估模型的性能,包括精确率(Precision)、召回率(Recall)和$F1$值($F1$)。其中,Precision表示真正的正样本与预测为正样本的比例;Recall表示被正确识别的正样本与总正样本的比例;$F1$表示精确率与召回率之间的权衡。

表2 实验数据集

Table 2 Experimental dataset

数据集	训练集	验证集	测试集	平均长度
IAC-V1	1 484	185	185	68.32
IAC-V2	3 715	464	464	55.82
Twitter(Ptáček)	8 497	1 062	1 062	18.09

实验参数设置:RoBERTa-Base模型具有12层,隐藏层维度为768;Batch Size设置为16;

Learning Rate 设置为 2e-5;Epoch 设置为 8;Max_seq_len 依据不同数据集进行调整,其中 IAC-V1 和 IAC-V2 设置为 128,Twitter(Ptáček)则设置为 50;优化器采用 Adam;Dropout 设置为 0.1。

实验硬件配置:Intel(R) Xeon(R) Silver 4310 2.10 GHz 12 核 CPU,32 GB 内存,NVIDIA GeForce RTX 4090 显卡。操作系统为 Ubuntu 22.04,深度学习框架采用 Pytorch 2.4.1。

3.2 性能对比实验和分析

为了验证 SIICE 模型在讽刺检测任务上的性能,本文与九种现有的基线模型进行了对比实验。这九种基线模型具体描述如下:

● 多维句内注意力循环网络(Multi-dimensionalIntra-Attention Recurrent Network,MIARN)[14]:通过多维句子内部注意力机制与 LSTM 编码器,有效建模词语间复杂的语义关系。

● 加权句段自注意力模型(Self-Attention of Weighted Snippets,SAWS)[6]:一种基于句子片段的加权自注意力机制,通过动态分配片段权重捕捉片段级语义不一致。

● BERT:一种基于 Transformer 架构的双向编码器的预训练语言模型。BERT 主要通过掩码语言模型(Masked Language Model)和下一句预测(Next Sentence Prediction)两个任务进行预训练。

● RoBERTa[30]:一种基于 BERT 改进的预训练模型,通过扩大训练数据、引入动态掩码及延长序列长度来捕获上下文语义信息。

● 多头不一致感知注意力网络(Multi-head-Incongruity Aware Attention Network,MIAN)[1]:结合多头自匹配网络与多头共匹配网络,分别建模目标文本内部及目标文本与上下文之间的语义不一致。

● 情感依存图卷积网络(Affective Dependency Graph Convolutional Network,ADGCN)[11]:将情感图与依赖图相结合,采用多层图卷积网络交互建模情感与句法特征,并引入注意力机制从上下文表示中捕获情感依赖图导向的特征。

● 常识情感依存图卷积网络(Commonsense Sentiment Dependency Graph Convolutional Network,CSDGCN)[9]:利用 COMET 模型生成常识知识,并构建常识增强的情感图与常识替换的依赖图,通过图卷积网络捕捉文本中的语义矛盾。

● 情感不一致常识推理框架(Emotional Incongruity Commonsense Reasoning,EICR)[20]:利用检索增强的 GPT-4o 补全常识,结合优化依存图和情感不一致推理,通过层次化图注意力提取关键语义特征。

● SIICE-BERT:SIICE 模型的一个版本,将文本编码器由 RoBERTa 替换为 BERT,以对比不同预训练模型的性能。

本文提出的 SIICE 模型与九个基线模型的实验结果如表 3 所示。(表 3 中性能最好的指标用粗体标出,Precision、Recall、*F*1 三个指标数值均越大越好)。

表 3 不同讽刺检测模型的对比实验结果

Table 3 Results of different sarcasm detection models

Models	Twitter(Ptáček)			IAC-V1			IAC-V2		
	Precision	Recall	*F*1	Precision	Recall	*F*1	Precision	Recall	*F*1
MIARN	0.850 2	0.842 7	0.845 9	0.638 8	0.637 1	0.631 8	0.729 2	0.729 3	0.727 5
SAWS	0.831 8	0.831 2	0.831 4	0.662 2	0.656 5	0.656 0	0.735 2	0.734 0	0.734 3
BERT	0.839 9	0.846 5	0.842 4	0.649 5	0.691 8	0.670 0	0.751 1	0.766 3	0.753 6
RoBERTa	0.846 3	0.865 3	0.853 0	0.667 6	0.692 8	0.678 7	0.755 8	0.767 2	0.761 4
MIAN	0.854 3	0.870 4	0.860 9	0.675 0	0.712 7	0.684 0	0.764 3	0.770 7	0.766 8
ADGCN	0.841 1	0.848 7	0.844 9	0.661 4	0.686 4	0.673 7	0.761 0	0.760 9	0.761 0
CSDGCN	0.863 7	0.860 3	0.863 6	**0.689 5**	0.676 5	0.688 8	0.767 5	0.763 0	0.766 5
EICR	0.861 3	0.857 9	0.860 0	0.670 3	0.680 3	0.674 5	0.754 2	0.767 2	0.760 6
SIICE-BERT	0.864 2	0.870 2	0.867 2	0.671 7	0.718 9	0.694 5	0.751 5	0.790 5	0.770 0
SIICE	**0.868 0**	**0.872 9**	**0.870 3**	0.681 8	**0.729 7**	**0.704 9**	**0.767 6**	**0.806 3**	**0.783 5**

注:粗体表示性能最好的指标。

由表3可见,本文提出的SIICE模型在三个基准数据集上整体表现均优于其他九个基线模型,$F1$值分别达到0.870 3(Twitter(Ptáček)),0.704 9(IAC-V1)和0.783 5(IAC-V2)。具体分析如下:

首先,在Twitter(Ptáček)数据集中,MIARN和SAWS模型表现相对较差,表明单纯依赖句内特征的建模方法难以有效捕捉复杂的语义差异。与BERT模型相比,RoBERTa模型的$F1$值提升了0.010 6,这表明RoBERTa通过优化的预训练策略(如动态掩码和更大规模的训练数据),在抽取文本语义信息方面相较于原始BERT模型具备明显的优势。此外,与RoBERTa、MIAN、ADGCN、CSDGCN及EICR模型相比,SIICE模型的$F1$值分别提升了0.017 3、0.009 4、0.025 4、0.006 7、0.010 3。这一结果表明,相较于仅关注文本内部或单独考虑文本与单一类型外部知识之间语义不一致的方法,SIICE模型通过同时建模文本与两种外部知识之间的语义不一致,更有效地提升了讽刺检测性能。类似地,SIICE-BERT模型的$F1$值也超过了其他基线模型,这与SIICE模型的实验结论相一致,表明不同预训练语言模型的选择对SIICE模型架构的整体性能影响较小。

在IAC-V1和IAC-V2数据集中,SIICE模型的$F1$值均优于九个基线模型。这一结果表明,在长文本复杂语境下,同时捕捉文本与常识知识、文本与情绪知识之间的语义不一致,有效提高了模型对复杂讽刺的识别能力。相比之下,仅关注单一类型的语义关联在复杂对话场景中表现出明显的局限性。值得注意的是,在IAC-V1数据集中,虽然CSDGCN模型在精确率上略高于SIICE模型(高出0.007 7),但召回率却显著偏低,整体性能不如SIICE模型。这一现象可能源于CSDGCN模型倾向于捕捉明显的表层特征,而在处理语义边界模糊或隐性讽刺线索时泛化能力不足。

3.3 消融实验

为系统评估SIICE模型中各模块的有效性,本文进行了消融实验(结果见表4)。其中,w/o Comet、w/o VAD、w/o Text分别表示移除了常识知识、情绪知识、文本语义不一致感知模块。

在Twitter(Ptáček)数据集中,w/o Text模型的性能下降最为明显,$F1$值降低了0.013 4,这说明短文本讽刺主要依赖词语之间的语义冲突,模型较容易捕捉其内部的语义不一致,凸显了文本自身语义不一致在讽刺检测中的关键作用;而w/o Comet模型和w/o VAD模型的$F1$值分别仅降低了0.003 8和0.008 8,说明单独探讨文本与常识知识或文本与情绪知识之间的语义不一致难以全面弥补短文本语境中缺失的外部知识;当同时考虑文本与两种外部知识的语义不一致时,SIICE模型性能达到最佳。

表4 消融实验结果

Table 4 Results of the ablation experiment

Method	Twitter (Ptácek)			IAC-V1			IAC-V2		
	Precision	Recall	$F1$	Precision	Recall	$F1$	Precision	Recall	$F1$
w/o Comet	0.861 2	0.873 3	0.866 5	0.675 3	0.697 2	0.686 1	0.756 9	0.783 8	0.766 8
w/o VAD	0.842 0	**0.883 4**	0.861 5	0.658 2	0.708 1	0.682 2	0.766 9	0.780 1	0.773 5
w/o Text	0.850 8	0.863 7	0.856 9	0.666 3	0.710 1	0.685 9	0.756 0	0.802 4	0.761 7
SIICE	**0.868 0**	0.872 9	**0.870 3**	**0.681 8**	**0.729 7**	**0.704 9**	**0.767 6**	**0.806 3**	**0.783 5**

注:粗体表示性能最好的指标。

在IAC-V1数据集中,移除三个模块后模型性能均显著降低,$F1$值分别下降了0.018 8(w/o Comet)、0.022 7(w/o VAD)及0.019(w/o Text)。这一结果表明,在处理长文本时,三种语义不一致信息对讽刺检测任务存在明显的互补作用。相对而言,在IAC-V2数据集中,w/o Text模型性能下降最明显($F1$值下降了0.021 8),凸显了文本语义不一致信息在长文本讽刺识别中的关键作用;而w/o Comet模型和w/o VAD模型性能下降幅度相对较小,说明在长文本中讽刺现象通常隐匿于更为复杂的真实语境中,仅依赖单一的情绪知识或常识知识语义不一致信息难以有效捕捉其中的语义冲突。同时融合文本与两种外部知识之间的语义不一致信息,有助于更全面地揭示长文本复杂讽刺表达所蕴含的语义冲突,从而提升讽刺识别的

性能。

3.4 案例分析

为阐明本文融合常识与情绪知识以捕捉语义不一致的作用机制,本文分别从三个数据集中各选取了一个讽刺案例进行分析(见表5)。这些案例共同揭示了讽刺是常识与情绪双重作用下的复杂语言现象,单一维度的外部知识不足以完全揭示讽刺的复杂性。例如句子"Oh great, another Monday. Just what I needed!",首先从常识维度分析,讽刺产生于句子字面含义与社会文化常识之间的语义冲突。社会文化常识将"周一"这一概念与工作压力、疲惫等负面状态关联,而句子却采用了"great""just what I needed"等积极词汇进行描述。这种字面措辞与常识知识之间的不一致构成了讽刺;其次,从情绪维度分析,句子体现出典型的情绪反转特征,即表面表达的积极情绪与"another Monday"所隐含的真实负面情绪(如沮丧、厌烦)之间存在冲突。这种表层情绪与真实意图之间的不一致是讽刺的关键来源之一。其余两个案例也呈现出类似的不一致特征,表现为文本字面含义或表面情绪与社会常识、真实情绪意图之间的语义冲突。因此,单独探讨文本与常识知识或文本与情绪知识之间的语义不一致存在局限性,同时考虑文本与两种外部知识之间的语义不一致,模型才能更准确地理解和识别讽刺,从而提升讽刺检测的性能。

表5 三个数据集中的部分案例

Table 5 Some cases from the three datasets

Prediction:sarcasm Real label:sarcasm Oh great, another Monday. Just what I needed!
Prediction:sarcasm Real label:sarcasm That awesome feeling when you try to do something nice and that person doesn't even care.
Prediction:sarcasm Real label:sarcasm I love these study hours in junker!

4 总结与展望

本文针对文本与常识知识、文本与情绪知识之间存在的语义不一致现象,提出了一种基于常识知识和情绪知识语义不一致的讽刺检测方法(SIICE)。SIICE模型包含四个模块:文本语义不一致感知模块、常识知识语义不一致感知模块、情绪知识语义不一致感知模块和融合预测模块。与现有讽刺检测模型不同,SIICE模型利用注意力机制对文本语义、常识知识和情绪知识相互之间的不一致进行建模,融合三种不一致信息并与原始文本语义信息拼接后进行讽刺检测。在三个公开数据集上的实验结果表明,SIICE模型性能明显优于基线模型。消融实验进一步验证了同时融合文本与两种外部知识之间的语义不一致信息能够有效提升模型性能。

未来研究将进一步扩展至多模态领域,探索图像、语音等跨模态语义不一致特征,以提升讽刺检测的准确性与鲁棒性。

参考文献:

[1] Guan X, Cao J X, Zhang H, *et al*. MIAN: Multi-head Incongruity Aware Attention Network with Transfer Learning for Sarcasm Detection[J]. *Expert Syst Appl*, 2025, **263**: 125702. DOI: 10.1016/j.eswa.2024.125702.

[2] Oprea S, Magdy W. ISarcasm: A Dataset of Intended Sarcasm[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: ACL, 2020: 1279–1289. DOI: 10.18653/v1/2020.acl-main.118.

[3] Ilievski F, Oltramari A, Ma K X, *et al*. Dimensions of Commonsense Knowledge[J]. *Knowl Based Syst*, 2021, **229**: 107347. DOI: 10.1016/j.knosys.2021.107347.

[4] Tu G, Liang B, Qin B, *et al*. An Empirical Study on Multiple Knowledge from ChatGPT for Emotion Recognition in Conversations[C]//Findings of the Association for Computational Linguistics: EMNLP 2023. Stroudsburg, PA, USA: ACL, 2023: 12160–12173. DOI: 10.18653/v1/2023.findings-emnlp.813.

[5] Xiong T, Zhang P R, Zhu H B, *et al*. Sarcasm Detection with Self-matching Networks and Low-rank Bilinear Pooling[C]//The World Wide Web Conference. New York: ACM, 2019: 2115–2124. DOI: 10.1145/3308558.3313735.

[6] Pan H L, Lin Z, Fu P, *et al*. Modeling the Incongruity Between Sentence Snippets for Sarcasm Detection[C]//Proceedings of the 24th European Conference on Artificial Intelligence (ECAI). Amsterdam, Netherlands: IOS Press BV, 2020: 2132–2139. DOI: 10.3233/FAIA200337.

[7] Li J N, Pan H L, Lin Z, *et al*. Sarcasm Detection with Commonsense Knowledge[J]. *IEEE/ACM Trans Audio Speech Lang Process*, 2021, **29**: 3192–3201. DOI: 10.1109/TASLP.2021.3120601.

[8] Chen W Q, Lin F Q, Li G W, *et al*. Commonsense-aware Sarcasm Detection with Heterogeneous Graph Attention Network[C]//2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC). New York: IEEE, 2022: 2181−2188. DOI: 10.1109/SMC53654.2022.9945145.

[9] Yu Z, Jin D, Wang X B, *et al*. Commonsense Knowledge Enhanced Sentiment Dependency Graph for Sarcasm Detection[C]//Proceedings of the Thirty−Second International Joint Conference on Artificial Intelligence. Macao: IJCAI, 2023: 2423−2431. DOI: 10.24963/ijcai.2023/269.

[10] Babanejad N, Davoudi H, An A J, *et al*. Affective and Contextual Embedding for Sarcasm Detection[C]//Proceedings of the 28th International Conference on Computational Linguistics. Barcelona: ICCL, 2020: 225−243. DOI: 10.18653/v1/2020.coling-main.20.

[11] Lou C W, 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. New York: ACM, 2021: 1844−1849. DOI: 10.1145/3404835.3463061.

[12] Wang X, Dong Y, Jin D, *et al*. Augmenting Affective Dependency Graph via Iterative Incongruity Graph Learning for Sarcasm Detection[C]//Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). Washington, USA: AAAI Press, 2023, **37**(4): 4702−4710. DOI: 10.1609/aaai.v37i4.25594.

[13] Young T, Pandelea V, Poria S, *et al*. Dialogue Systems with Audio Context[J]. *Neurocomputing*, 2020, **388**: 102−109. DOI: 10.1016/j.neucom.2019.12.126.

[14] Tay Y, Luu A T, Hui S C, *et al*. Reasoning with Sarcasm by Reading In-between[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg, PA, USA: ACL, 2018: 1010−1020. DOI: 10.18653/v1/p18-1093.

[15] Poria S, Cambria E, Hazarika D, *et al*. A Deeper Look into Sarcastic Tweets Using Deep Convolutional Neural Networks[C]//Proceedings of the 26th International Conference on Computational Linguistics (COLING). Osaka, Japan: The COLING 2016 Organizing Committee, 2016: 1601−1612. DOI: 10.48550/arXiv.1610.08815.

[16] 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 (COLING). Santa Fe, USA: Association for Computational Linguistics, 2018: 1837−1848.

[17] Speer R, Chin J, Havasi C. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge[C]//Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). San Francisco, USA: AAAI Press, 2017, **31**(1): 4444−4451. DOI: 10.1609/aaai.v31i1.11164.

[18] Sap M, Le B R, Allaway E, *et al*. ATOMIC: An Atlas of Machine Commonsense for If-Then Reasoning[C]//Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). Honolulu, USA. AAAI Press, 2019, **33**(1): 3027−3035. DOI: 10.1609/aaai.v33i01.33013027.

[19] Bosselut A, Rashkin H, Sap M, *et al*. COMET: Commonsense Transformers for Automatic Knowledge Graph Construction[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: ACL, 2019: 4762−4779. DOI: 10.18653/v1/p19-1470.

[20] Qiu Z, Yu J, Zhang Y, *et al*. Detecting Emotional Incongruity of Sarcasm by Commonsense Reasoning[C]//Proceedings of the 31st International Conference on Computational Linguistics (COLING). Abu Dhabi, UAE: Association for Computational Linguistics, 2025: 9062−9073.

[21] Russell J A, Mehrabian A. Evidence for a Three-factor Theory of Emotions[J]. *J Res Pers*, 1977, **11**(3): 273−294. DOI: 10.1016/0092-6566(77)90037-X.

[22] Savini E, Caragea C. A Multi-task Learning Approach to Sarcasm Detection (Student Abstract)[J]. *Proc AAAI Conf Artif Intell*, 2020, **34**(10): 13907−13908. DOI: 10.1609/aaai.v34i10.7226.

[23] Chauhan D S, Dhanush S R, Ekbal A, *et al*. Sentiment and Emotion Help Sarcasm? A Multi-task Learning Framework for Multi-modal Sarcasm, Sentiment and Emotion Analysis[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: ACL, 2020: 4351−4360. DOI: 10.18653/v1/2020.acl-main.401.

[24] Agrawal A, An A J, Papagelis M. Leveraging Transitions of Emotions for Sarcasm Detection[C]//Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2020: 1505−1508. DOI: 10.1145/3397271.3401183.

[25] Chakrabarty T, Ghosh D, Muresan S, *et al*. R^3: Reverse, Retrieve, and Rank for Sarcasm Generation with Commonsense Knowledge[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: ACL, 2020: 7976−7986. DOI: 10.18653/v1/2020.acl-main.711.

[26] 赖金水, 万中英, 曾雪强. 基于情感轮和多任务卷积神经网络的图像情感分布学习[J]. 江西师范大学学报

(自然科学版), 2022, **46**(4): 363−371. DOI: 10.16357/j.cnki.issn1000-5862.2022.04.06.
Lai J S, Wan Z Y, Zeng X Q. The Image Emotion Distribution Learning Based on Emotion Wheel and Multi-task Convolutional Neural Network[J]. *J Jiangxi Norm Univ Nat Sci Ed*, 2022, **46**(4): 363−371. DOI: 10.16357/j.cnki.issn1000-5862.2022.04.06.

[27] 王耀琦, 万中英, 曾雪强, 等. 融合VAD情绪知识的文本情感分布标签增强方法[J]. 清华大学学报(自然科学版), 2024, **64**(5): 789−800. DOI: 10.16511/j.cnki.qhdxxb.2023.26.063.
Wang Y Q, Wan Z Y, Zeng X Q, *et al*. Valence-arousal-dominance Emotion Knowledge-based Text Emotion Distribution Label Enhancement Method[J]. *J Tsinghua Univ Sci Technol*, 2024, **64**(5): 789−800. DOI: 10.16511/j.cnki.qhdxxb.2023.26.063.

[28] Walker M A, Anand P, Tree J E F, *et al*. A Corpus for Research on Deliberation and Debate[C]//Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC). Istanbul, Turkey: European Language Resources Association, 2012, **12**: 812−817.

[29] Ptacek T, Habernal I, Hong J. Sarcasm Detection on Czech and English Twitter[C]//Proceedings of the 25th International Conference on Computational Linguistics (COLING). Dublin, Ireland: Dublin City University and Association for Computational Linguistics, 2014: 213−223.

[30] Semary N A, Ahmed W, Amin K, *et al*. Improving Sentiment Classification Using a RoBERTa-based Hybrid Model[J]. *Front Hum Neurosci*, 2023, **17**: 1292010. DOI: 10.3389/fnhum.2023.1292010.