DOI:10.3969/j.issn.2096-8248.2024.04.010

公共事件下强弱不良模因识别方法*

仲兆满[1,2],杜家云[1]

(1.江苏海洋大学 计算机工程学院,江苏 连云港　222005;2.江苏省海洋资源开发研究院,江苏 连云港　222005)

摘　要:针对目前围绕不良模因研究较少的问题,构建公共事件下不良模因识别模型。以 2020 年“7·5 杭州女子失踪案”这一公共事件社交媒体数据为例,揭示不良模因的内涵,根据不良模因的类型和特性,提出 isMeme 特征,构建识别模型,对模型进行训练和评估,找出最优识别模型。实验结果表明,isMeme 特征能够有效实现不良模因的识别。识别模型中,SVM 模型表现最好,准确率达到 95%。该研究并未考虑不良模因出现早期的识别问题,后期可进一步分析其早期特征,在数据量更少的情况下实现有效识别。

关键词:公共事件;不良模因;isMeme 特征;识别模型

中图分类号:C912.63;G206　　**文献标志码**:A　　**文章编号**:2096-8248(2024)04-0075-08

引用格式:仲兆满,杜家云.公共事件下强弱不良模因识别方法[J].江苏海洋大学学报(自然科学版),2024,33(4):75-82.

Method for Identifying Strong and Weak Bad Memes in Public Events

ZHONG Zhaoman[1,2], DU Jiayun[1]

(1. School of Computer Engineering, Jiangsu Ocean University, Lianyungang 222005, China;
2. Jiangsu Institute of Marine Resources Development, Lianyungang 222005, China)

Abstract: To address the current lack of research on bad memes, a model for identifying bad memes in public events is constructed. Taking the social media data of the “July 5th Hangzhou Women's Disappearance Case” in 2020 as an example, this study reveals the connotation of bad memes. Based on the types and characteristics of bad memes, isMeme features are proposed, and a recognition model is constructed. The model is trained and evaluated to find the optimal recognition model. The experimental results indicate that isMeme features can effectively identify bad memes. Among the recognition models, the SVM model performs the best with an accuracy of 95%. This research does not consider the early identification of bad memes, and further research can analyze their early features to achieve effective identification with less data.

Key words: public events; bad memes; isMeme feature; recognition model

0　引言

互联网技术的发展加速了信息的传播。公共事件发生后,各种信息在社交媒体上迅速传播。模因(meme)具有自我复制、演化和传播的特性,在生物学和文化学上都广为使用,公共事件衍生的不良模因常常给舆情控制带来挑战。例如 2020 年的“7·5 杭州女子失踪案”网络上出现的“化粪池警告”等不良模因,消解了公共事件的严肃性,对当事人造成了

* **收稿日期**:2024-01-08;**修订日期**:2024-03-11
基金项目:国家自然科学基金资助项目(72174079)
作者简介:仲兆满(1977—),男,江苏连云港人,教授,博士,研究方向为互联网舆情大数据分析及管控,(E-mail)zhongzhaoman@163.com。

二次伤害,严重危害了网络生态环境,造成了恶劣的社会影响。因此,实现不良模因的有效识别具有重要的现实意义。

已有不少学者关注各类不良模因的识别。Shang 等[1]设计捕捉在线社交媒体中冒犯性模因隐含的类比以实现该类模因的识别。Nayak 等[2]演示了各种机器学习模型来自动检测互联网中的仇恨模因。Shang 等[3]开发了一种知识丰富的图神经网络,利用常识来有效检测在线社交媒体上的攻击性模因。学术界对不同类型的模因采用了不同的方法进行识别,却少有学者关注如何识别公共事件衍生的不良模因。

基于此,本研究收集微博、抖音和微信平台上的不良模因数据,给出公共事件下不良模因的定义,针对不同类型不良模因的识别,提出 isMeme 特征,结合 SVM(support vector machines,支持向量机)、DT(decision tree,决策树)、RF(random forest,随机森林)、LR(logistic regression,逻辑回归)、NB(naive bayes,朴素贝叶斯)、AdaBoost(adaptive boosting,自适应增强)和 KNN(K-nearest neighbor,K 最近邻)等模型实现对不良模因的识别,并对模型进行训练和评估,找出最优模型。识别不良模因,有利于网络生态环境的治理,帮助舆情管理部门了解公共事件舆情走向,防止次生舆情的发生,为舆论引导和危机治理提供方法支持。本文的主要贡献为:① 结合已有研究揭示公共事件下不良模因的内涵;② 提出 isMeme 特征,根据不同类型不良模因的特点,设计不同的特征算法,并结合其他基础特征实现对不良模因的有效识别。

1 相关工作

1.1 互联网中的模因

模因一词起源于 Richard Dawkins 的著作 *The Selfish Gene*,它是一种能够通过模仿而被复制的信息单位[4]。任何可以通过模仿被复制传播出去的信息,均为模因,模因可以看作文化传播的“基因”。模因在互联网时代之前就已经被广泛记录,而在线社交网络的发展加速、丰富了其产生与生长历程。任何出现在互联网上并被无数传播参与者模仿、重混合迅速扩散而产生无数衍生物的人工信息均为互联网模因[5],其类别包括且不限于短视频、话题标签、表情包、谣言、虚假信息和网络热梗等。

本文所研究的公共事件下的模因与“梗”、网络流行语等类似。梗的诞生发展分为语音变异和词汇变异两大源头,造梗者借用某些词句或者事物充当语料,玩梗者结合个人对造梗者的输出进行理解和再造[6]。李欣等[7]也发现作为亚文化的网络流行语,使用了很多仿拟和隐喻的修辞手法,具有一种“黑色幽默”。不良模因更贴近于“烂梗”。谢卓[8]以俄乌冲突下我国的网络空间中出现的许多如“收留乌克兰美女”之类烂梗为例,探讨了“玩梗”危机带出的网民媒介素养议题。“收留乌克兰美女”这一烂梗就是由公共事件“2022 年俄乌战争”衍生的不良模因,该不良模因的传播,不仅漠视了战争的严肃性,还对我国国际形象造成了损害。林爱珺[6]认为,梗文化依赖的是将逻辑推至极致带来的荒诞,并借助于这一荒诞性去完成一次笑点的制造。随着时代发展,网络玩梗文化兴起,梗文化的流行一定程度上为社交文化带来了新气象,但是应当警惕玩梗带来的传播迷思、娱乐至死和价值取向偏差。

1.2 公共事件下的互联网模因分析及识别

公共事件在网络中发酵、传播,常常伴随着模因的产生与发展。诸多学者探讨了不同类型公共事件产生的模因,其中,政治事件下的模因受到研究人员的关注。Galipeau[9]探讨了社交媒体上的政治模因对公民政治态度的影响;Mcloughlin 等[10]从创造者、参与度以及模因包含的政治信息 3 个方面分析了模因在政治运动中的作用;Lukács[11]专注于匈牙利所谓的 O1G 激进主义,证明模因在政治激进主义中的作用取决于它们产生的特定社会政治背景。

学术界在识别谣言、虚假信息等方面付出了巨大努力。Tian 等[12]考虑到谣言检测中的广泛分散结构,提出了一种新颖的双向图模型,通过自上而下和自下而上的谣言传播方式探索谣言在传播和散布两个方面的关键特征。Munyole 等[13]注意到长期被忽略的双重情感特征(出版商情感和社会情感),提出了一种基于深度归一化注意力的机制,用于丰富提取双重情绪特征,以及用于分类的自适应遗传权重更新-随机森林模型。相比之下,对模因尤其是不良模因进行建模并未受到太多关注。

国内学者对模因的研究一方面从传播学、语言学入手探讨其社会影响、大众心理和语法构词等,另一方面主要搭建模型以实现对谣言、虚假信息等广义模因的检测。国外学术界除此之外,也结合政治事件、公共卫生事件等探讨模因,但是公共事件下的不良模因仍少有文献提及。基于此,文章揭示了公

共事件下不良模因的内涵,提出了 isMeme 特征,并结合 SVM,DT,RF,LR,NB,AdaBoost,KNN 等模型实现对不良模因识别,并对模型进行训练和评估,找出最优预测模型。

2　不良模因识别模型

2.1　词典构建

本文采用基于统计、规则和手动等方法构建两性词典 Gender,高频词词典 HFreq,实体词典 Victim,Hero,Evil 和 Other。

Gender 词典包含与女性或两性相关的字词,目前并没有类似的公开词典可以直接使用,故基于具体事件的数据设定正则表达式,找出数据中所有带有“女”的字词添加到词典中。由于数据量有限,同时结合网上搜索、线下查阅等手动方式搜集符合条件的字词。HFreq 词典中包含具体公共事件不良模因数据高频词,基于统计的方法构建该词典。数据预处理后,利用 python 得到数据的词频统计结果,选择词频排名前 N 的词添加到 HFreq 中,$N>0$ 且 N 的大小与数据集大小正相关。NHFreq 词典中包含具体公共事件非不良模因数据高频词,基于统计的方法构建该词典。数据预处理后,利用 python 得到数据的词频统计结果,选择词频排名前 M 的词添加至 NHFreq 中,$M>0$ 且 M 的大小与数据集大小正相关。实体词典为 Victim,Hero,Evil 和 Other,分别包含事件中的受害者、英雄、恶棍和其他实体。基于 HFreq 和 NHFreq,手动选取两高频词词典中的对应实体添加到对应词典中,该词典的构建很大程度上依赖于模因“通过模仿而复制”的特性,即认为不良模因数据集中的数据具有高度相似性。

2.2　相关定义

互联网模因常通过拼贴、戏仿等手段产生,有“言在此而意在彼”的特点,含义有时隐晦,甚至在较短的语句中充满了语法错误,与网络梗、网络流行语等在定义上相似。Yao[14] 发现网络流行语中的大多数都不是凭空出现的,而是现存单词或表达方式的变体,它们继承、扩展或有时推翻了原来的含义。互联网模因种类繁多,但也存在共性。Paciello 等[15] 认为模因最典型的特征就是具有幽默意味,但 Hofer 等[16] 指出,模因的幽默通常建立在刻板印象之上,其中潜在的更具破坏性的捏造是显而易见且令人不安的。

表 1 和表 2 分别是近年来部分公共事件衍生的正常模因和不良模因。

表 1　公共事件衍生的正常模因

Table 1　Normal memes derived from public events

事件名称	模因	释义
“3·14 拉萨打砸抢烧暴力犯罪事件”	做人不能太 CNN	2008 年,境内外“藏独”势力勾结制造拉萨市区暴乱,一些外国媒体报道时肆意歪曲事实。此梗清楚地表明中国人民对西方媒体歪曲报道的反感
“7·23 甬温线特别重大铁路交通事故”	不管你信不信,反正我信了	2011 年,7·23 甬温线特别重大铁路交通事故新闻发布会上,铁道部新闻发言人称“这只能说是生命的奇迹”,还说出“至于你信不信,我反正信了”等话,引起广大网民的不满
“8·26 包茂高速特大交通事故”	表哥	2012 年,陕西省安监局原局长面带微笑出现在特大交通事故现场,随后,其在不同场合佩戴多块名表的照片被曝光,被网友称为“表哥”

表 2　公共事件衍生的不良模因

Table 2　Bad memes derived from public events

事件名称	模因	释义
“2022 年俄乌冲突”	收留乌克兰美女	俄乌战争中,乌克兰产生大量难民,网络上出现关于“收留乌克兰美女”的言论
“6·10 唐山烧烤店打人事件”	唐山烧烤	2022 年唐山某烧烤店发生恶性打人事件,且殴打之前有骚扰、侮辱女性的严重情节。事件曝光后,网络上出现“带你去唐山吃烧烤”等不当言论
“7·5 杭州女子失踪案”	① 化粪池警告;② 两吨水解决;③ 交水费;④ 同款绞肉机;⑤ 感谢老公不杀之恩	2020 年某女子被丈夫杀害,尸体被丢进化粪池。许多网友推测凶手将尸体用绞肉机粉碎后用了两吨水冲进化粪池,并产生高额水费
“3·10 埃塞俄比亚航班坠毁事故”	千里送炮	2019 年一架坠毁客机上有一名与男友相约赴非洲的女性乘客。事发后,许多网友在该女性社交账号下发表不当言论

结合表 1、表 2,公共事件衍生的正常模因常反映公众对事件的看法,或暗含对事件相关人物、事件本身的讽刺,或传达公众面对公共事件的心态。而不良模因往往具有贬低幽默,即通过诋毁、贬损或贬低给定目标(如个人、社会团体、政治意识形态、物质财产)来引起娱乐的幽默,严重危害了网络生态环境,也常引起次生舆情,加大了政府相关部门舆情管控的难度。仲兆满等[17]指出,研究互联网公共事件相关信息能提升管控效能。有效识别不良模因对把握网络舆情动态至关重要。

综上,给出公共事件下各类模因的定义。

定义 1 模因,公共事件情境下,指伴随公共事件 $E=\{\mathrm{EE},\mathrm{NEE}\}$ 的发生,在社交网络被反复引用、不断演化且含隐晦意义、有幽默意味的词、短语、句子或特定的句子模式。其中 EE(Entity Elements,实体元素),NEE(Non-entity Elements,非实体元素)是公共事件 E 的构成元素。

定义一个函数 G,G 结合 E 中的元素映射对事件 E 的看法或态度,形成正常模因集合 $\mathrm{NMeme}=\{N_1,N_2,\cdots,N_n\}$。

定义 2 正常模因,公共事件情境下,指借用与公共事件相关的元素,表达对事件的看法或态度的模因。形式化表达为:$\mathrm{NMeme}=\{G(\mathrm{EE})+G(\mathrm{NEE})\}$。

相较于正常模因,不良模因常涉及对女性的刻板印象。女性比男性更加容易遭受网络暴力。Lawless 等[18]发现针对女性的笑话被认为比针对男性的笑话更令人反感、更性别歧视。国内外许多文献从社会学、心理学等不同角度探讨了对女性的性别歧视,幽默可以作为一种微妙的方式来表达偏见,有学者将对女性的性别歧视与幽默联系起来。Aprianti 等[19]研究发现如果不对性别歧视幽默进行预防和妥善处理,会发展成非口头的暴力。Mallya 等[20]发现一些模因通过性别歧视的幽默传达了羞辱身体的信息,主要针对肥胖/有色人种/跨性别女性。关于性别歧视的幽默虽然披了一层笑话和娱乐的外衣,但仍不掩其对性别歧视容忍的本质。公共事件下的不良模因常借由公共事件,以幽默的口吻表达对女性的不当调侃。

用 dic(Gender)表示 Gender 词典,$E'=\{\mathrm{EE}\}$,函数 F 结合 E' 和 dic(Gender)中的元素生成不良模因集合 $\mathrm{BMeme}=\{B_1,B_2,\cdots,B_m\}$。

定义 3 不良模因,公共事件情境下,指借用与公共事件相关的实体,对他人(尤其是女性)表达恶意,具有攻击性或者涉及对两性关系不当调侃的模因,通常是词或短语。形式化表达为:$\mathrm{BMeme}=\{F(\mathrm{EE}+\mathrm{dic(Gender)})\}$。

定义 4 不良模因模板,通常是词语或者短语,是不良模因不断演化的模板。形式化表达为:$\mathrm{Template}=\{\mathrm{New(EE)},\mathrm{Phrase(EE)}\}$,根据不良模因涉及的实体,通过旧词改编、短语创造等方法得到。Template 集合包含词和短语两种不良模因模板。函数 New 基于实体元素实现旧词重构,生成词语类不良模因模板。函数 Phrase 将实体元素(名词)结合动词、形容词等形成短语类不良模因模板。

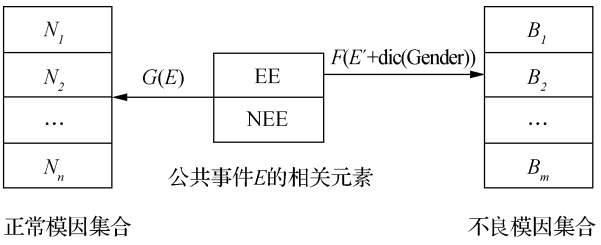

图 1 模因产生过程

Fig. 1 Meme generating process

定义 5 强不良模因,由不良模因模板和一些附加信息组成的语句。形式化表达为:$\mathrm{StrongBMeme}=\{\mathrm{Template}+\mathrm{text}\}$,$t\geqslant t_1$,text 常常与女性或者两性关系相关,$t$ 是社交媒体平台上不良模因发布的时间,t_1 是社交媒体上公共事件开始传播的时间。

定义 6 弱不良模因,由不良模因模板涉及的公共事件实体元素和一些附加信息组成的语句。形式化表达为:$\mathrm{WeakBMeme}=\{\mathrm{select(EE)}+\mathrm{text}\}$,$t_1\leqslant t\leqslant t_2$。其中,select(EE)生成 Template 中涉及的实体元素,text 常常与女性或者两性关系相关,t_2 是公共事件在社交媒体上结束传播的时间。通常情况下,不良模因模板为词语时,产生弱不良模因的概率较小,本文不做讨论。

2.3 数据来源

2020 年“7·5 杭州女子失踪案”衍生大量不良模因,具有代表性,收集该案件侦破后发布的官方微博下网友评论信息共计 6 923 条,去除仅有标点符号、表情符号以及与事件无关的,或者评论内容仅为“@用户 XX”等的无用信息后,经过清洗共有 5 473 条信息,其中正常信息 5 440 条,不良模因 33 条。为扩充不良模因数据量,在抖音、微信公众号共收集

不良模因 379 条,去重后余 87 条。最终构建的数据集中数据总量为 5 560 条,其中正常信息 5 440 条,不良模因 120 个。由于不良模因之间有较高的相似度,故本次实验构建的语料库中不良模因数量虽少,但实验结果具有客观性。

2.4　isMeme 特征构建

2.4.1　词类不良模因　词类不良模因在传播过程中形态稳定,少有突变。结合语料库中高频词词典 $HFreq=\{hfword_1,hfword_2,\cdots,hfword_N\}$ 和两性词典 $Gender=\{mword_1,mword_2,\cdots,mword_M\}$ 进行字符串匹配,匹配到则将文本 isMeme 特征值赋 1,否则为 0(见图 2)。

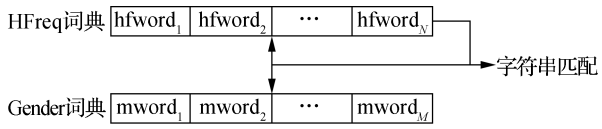

图 2　词类不良模因模板查找

Fig. 2　Search of bad memes template for words

2.4.2　短语类不良模因　短语类不良模因分为强不良模因和弱不良模因。强不良模因中包含完整的不良模因模板,传播范围更广。弱不良模因充分体现了社交媒体用户对模因的模仿与再创造能力,社交媒体用户通常结合模板中的实体元素进行再创造,相较于强不良模因其含义更为隐晦。对这两种不同的不良模因采取不同的方法进行识别,识别框架如图 3 所示。

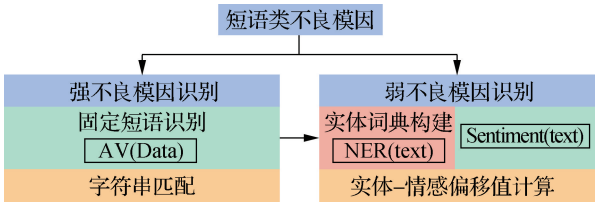

图 3　短语类不良模因识别框架

Fig. 3　Framework for identifying bad memes of phrases

(1) 强不良模因识别。Accessor Variety (AV) 分数可以反映单词在文本中的相对频率和与其他单词的相对差异。通过比较不同单词的 AV 分数,可以识别出具有相似 AV 分数的单词组合,这些组合可能构成固定短语。通常,固定短语具有较高的 AV 分数,因为其在文本中具有相对较高的频率和与其他单词的相对差异。据 AV 分数计算数据集中的独立短语算法如算法 1。

算法 1:独立短语识别

输入:数据集中提取的所有词 $tokens=\{word_1,\cdots,word_N\}$,阈值 threshold

输出:独立短语集合 phrases()

```
phrases()
bigrams = zip(tokens[:-1],tokens[1:])//得到双词
trigrams=zip(tokens[:-2],tokens[1:-1],tokens[2:])//得到三词
CalculateAV(bigrams)
CalculateAV(trigrams)
If AV>threshold
   phrases.add(trigrams)
If bigrams not in phrases
   phrases.add(bigrams)
   return phrases
```

基于已有数据运行该算法,得到固定短语:化粪池警告、两吨水解决、交水费、同款绞肉机、感谢老公,与表 2 中已知不良模因高度吻合。得到独立短语后进行字符串匹配,匹配到则将文本 isMeme 特征值赋 1,否则为 0。

(2) 弱不良模因识别。首先,构建实体词典。弱不良模因以模板中的实体部分更加隐晦地表达真实含义。将公共事件涉及主体分为 4 类:Victim(受害者),Hero(英雄),Evil(恶棍)和 Other(其他实体)。前 3 类为角色实体,最后 1 类既包括角色实体也包括物体等类型的实体。计算不良模因语料库中的高频词如表 3 所示。

表 3　两类信息高频词

Table 3　Two types of information high-frequency words

不良模因高频词				正常信息高频词			
词语	词频	词语	词频	词语	词频	词语	词频
化粪池	35	感谢	6	辛苦	1079	可能	223
两吨水	29	之恩	5	警察	707	工作人员	223
老公	28	吵架	5	真的	450	什么	211
绞肉机	17	每天	5	没有	376	这种	199
警告	14	解决	5	这个	375	自己	183
水费	11	解决不了	5	这么	333	如果	176
不听话	10	一天	4	不是	332	媒体	176
以后	9	同款	4	可怕	308	警方	175
结婚	9	回去	4	死刑	292	那么	172
老婆	9	小心	4	怎么	291	不会	171
不敢	8	扔进	4	分尸	281	应该	161
今天	8	时候	4	法医	250	觉得	161
可能	7	现在	4	就是	238	女儿	150
一切	6	一句	3	知道	229	残忍	149
家里	6	下水道	3	一个	226	化粪池	142

选取表中的名词,根据公共事件的具体情况,分别构建 Victim,Hero,Evil 和 Other 各实体词典如表 4 所示。

表 4 选取案例各实体词典

Table 4 Individual entity dictionaries for the selected cases

实体词典	实体元素
Victim	老婆
Hero	警察
Evil	老公
Other	化粪池、水费、绞肉机、两吨水

其次,计算实体-情感偏移度。通常情况下,积极的文本用于描述英雄,消极的文本则描述受害者和恶棍,而倾向于中性的文本则映射到其他实体。将 Victim 和 Evil 中实体情感赋值为−1,Hero 中赋值为 1,Other 中赋值为 0(其他公共事件中的情感分布不一定完全与之相同)。根据表 1、表 2,不良模因一般包含与公共事件相关的实体,且实体实际情感常与应有情感发生偏移,如图 4 所示。基于此,提出实体-情感偏移度 Devia。

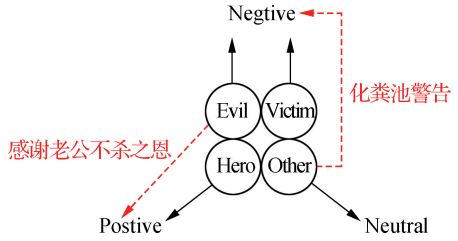

图 4 实体-情感偏移示意图

Fig. 4 Schematic diagram of entity-emotion migration

基于 BosonNLP 情感词典计算文本情感值,删除词典中不良模因涉及的实体词语。情感值为正,代表信息情感倾向为正;情感值为负,代表信息情感倾向为负。实体-情感偏移度具体算法见算法 2。

对于弱不良模因来说,isMeme=Devia,即当实体应有情感与实际情感发生偏移时,文本的 isMeme 值为 1,否则为 0。

算法 2:实体-情感偏移度

输入:数据集 Data=$\{text_1, text_2, \cdots, text_M\}$;各实体词典 Victim,Hero,Evil 和 Other;两性词典 Gender.

输出:各文本的实体-情感偏移度 Devia

Senti(Victim. txt)=Senti(Evil. txt)=−1

Senti(Hero. txt)=1

Senti(Other. txt)=0

Devia=[]

for i=1 to M do

Sentiment($text_i$)

NER($text_i$)//命名实体识别

If not entities or not has_entity_in(entities,Victim. txt,Hero. txt,Evil. txt,Other. txt)

Devia[i−1]=0

for entity in entities

if entity in Victim. txt or Evil. txt and sentiment($text_i$)!=−1

Devia[i−1]=1

elif entity in Hero. txtt and sentiment($text_i$)! =1

Devia[i−1]=1

elif entity in Other. txt and sentiment($text_i$)! =0

Devia[i−1]=1

if len(entities)>1 and has_entity_in(entities,Victim. txt,Hero. txt,Evil. txt,Other. txt) and has_word_in($text_i$,Gender. txt)

Devia[i−1]=1

return Devia

2.5 基础特征

Schlosberg[21]在 1954 年提出了激活度的概念,认为与愉悦情感相比,不愉悦的情感具有更高的激活度。对于“激活度”的量化,通常可以通过测量自主神经系统活动的变化来实现,比如心率、皮肤电导等。这些生理指标的改变可以反映人的情感状态。例如,当人们感到兴奋或紧张时,他们的心率可能会加快,皮肤电导可能会增加。由于社交媒体数据对生理指标的获取几乎是不可能的,故以文本中可以体现情感倾向、表达个人感情的词语作为量化激活度的指标。选取形容词、情感词、程度副词、语气词和否定词数量作为衡量激活度的标准,各实体情感-激活度坐标如图 5 所示。

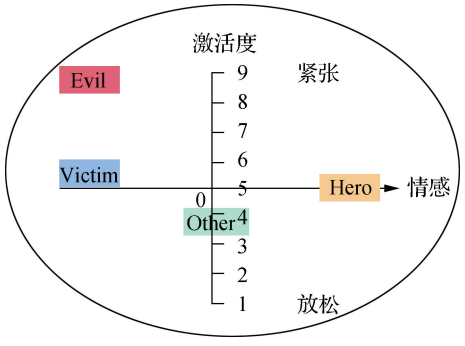

图 5 各实体情感-激活度坐标图

Fig. 5 Coordinates of emotion-activation of each entity

由于不良模因具有简洁性,且以公共事件实体词语(名词)为中心,同时选取名词数和句长作为基础特征,基础特征见表 5。

表 5　基础特征表

Table 5　Basic characteristics table

特征名称	特征表示	特征描述
激活度特征	Num_adj	形容词数
	Num_senti	情感词总数
	Num_adv	程度副词数
	Num_not	否定词数
	Num_tone	语气词数
其他基础特征	Num_n	名词数
	length	文本长度

3　实验分析

3.1　实验指标

使用 P,R 和 F_1 作为评价指标。其中,P 是准确率;R 是召回率;MCC 是可以测量二分类的分类性能的指标,它的取值范围为[−1,1],取值为 1 时表示对受试对象的完美预测,取值为 0 时表示预测的结果还不如随机预测的结果,−1 是指预测分类和实际分类完全不一致。具体计算方法如式(1)~(4)所示:

$$P=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}, \tag{1}$$

$$R=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \tag{2}$$

$$F_1=\frac{2\times P\times R}{P+R}, \tag{3}$$

$$\mathrm{MCC}=\frac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}。 \tag{4}$$

式中:TP 表示被模型预测为正类的正样本;TN 表示被模型预测为负类的负样本;FN 表示被模型预测为负类的正样本;FP 表示被模型预测为正类的负样本。

3.2　实验过程

不良模因的识别为二分类问题。实验数据量少,且机器方法可以更快地处理数据集,需要的计算资源也更少,更加适合处理小样本数据,故选用 SVM,DT,LR,NB,AdaBoost 和 KNN 6 种机器学习方法分别进行实验。DT,LR,NB,AdaBoost 和 KNN 等分类器分别通过在 scikit-learn 库中调用 sklearn. treer,LogisticRegression,MultinomialNB,AdaBoostClassifier 和 KNeighborsClassifier 模块实现。由于数据集正负样本分布失衡,对负样本进行下采样,重复实验 10 次,取各实验结果均值,实验结果如表 6 所示。

表 6　不同分类器下分类结果

Table 6　Classification results under different classifiers

分类器	准确率	召回率	F_1 值	MCC
SVM	**0.95**	**0.79**	**0.86**	**0.75**
DT	0.83	0.93	0.88	0.73
LR	0.85	0.85	0.85	0.70
NB	0.68	0.96	0.80	0.50
AdaBoost	0.55	0.84	0.56	0.27
KNN	0.82	0.96	0.89	0.74

SVM 分类器取得了最好的实验结果,分别选择其不同核函数得到分类结果如表 7 所示。可见当核函数选用 RBF 时,分类效果最佳,故后续实验分类器均选用 SVM 分类器,RBF 核函数。

表 7　SVM 不同核函数分类结果

Table 7　Classification results of different kernel functions of SVM

核函数	准确率	召回率	F_1 值	MCC
Linear	0.91	0.79	0.85	0.71
Poly	0.88	0.79	0.83	0.66
Sigmoid	0.91	0.79	0.84	0.70
RBF	**0.95**	**0.79**	**0.86**	**0.75**

不同特征组合下的实验结果见表 8。实验 A~C 特征组合均为基础特征,实验 A 仅基于名词数和句长特征进行分类,实验结果最差,只有不到 20% 的样本被正确分类。实验 B 使用了激活度特征,激活度特征能够体现文本从放松到紧张的情感状态,本文选取形容词、情感词、程度副词、语气词和否定词等词语数目以量化激活度,该组实验取得了较好的实验结果,相较于实验 A,实验效果有明显改进,验证了激活度特征的有效性。

实验 D 仅使用本文提出的 isMeme 特征。isMeme 特征对不同类型不良模因分情况做了细致讨论。本文选取案例为具有代表性的短语类不良模因模板,该特征对强不良模因进行字符串匹配,对于含义更为隐晦的弱不良模因,提出了实体-情感偏移度,该组实验取得了很好的实验效果,仅使用单个特征即达到 92% 的准确率,MCC 值也相对较高,验证了 isMeme 特征的有效性。实验 E 结合了基础特征和 isMeme 特征,此时分类器性能达到最佳。

表 8 不同特征集合的组合情况

Table 8 Combination of different feature sets

实验名称	特征组合	P	R	F_1	MCC
A	其他基础特征	0.58	0.89	0.70	0.19
B	**激活度特征**	**0.67**	**0.86**	**0.75**	**0.39**
C	基础特征	0.70	0.85	0.77	0.46
D	**isMeme 特征**	**0.92**	**0.79**	**0.85**	**0.70**
E	**基础特征+isMeme 特征**	**0.95**	**0.79**	**0.86**	**0.75**

4 结论

本文聚焦公共事件下的不良模因识别问题,分析社交媒体上不良模因与“梗”的联系与区别,概括模因的定义,对比正常模因揭示不良模因的内涵。在现有的激活度特征和其他基础特征的基础上,引入了 isMeme 特征,其对不同类型的不良模因采取不同的赋值方法,从而实现了基于 isMeme 和基础特征的不良模因识别模型,并通过实验验证本文所提出模型的有效性。未来会进一步研究不良模因在早期传播阶段的特点,结合更多有效的相关特征如用户特征和传播特征等,实现不良模因全阶段、多方面的有效识别。

参考文献:

[1] SHANG Lanyu, ZHANG Yang, ZHA Yuheng, et al. AOMD: an analogy-aware approach to offensive meme detection on social media[J]. Information Processing & Management, 2021, 58(5): 102664.

[2] NAYAK A, AGRAWAL A. Detection of hate speech in social media memes: a comparative analysis[C]// Third International Conference on Intelligent Computing Instrumentation and Control Technologies (ICICICT), 2022: 1179-1185.

[3] SHANG Lanyu, YOUN C, ZHA Yuheng, et al. KnowMeme: a knowledge-enriched graph neural network solution to offensive meme detection[C]//IEEE 17th International Conference on eScience (eScience), 2021: 186-195.

[4] DAWKINS R. The Selfish Gene[M]. UK: Oxford University Press, 2006.

[5] DYNEL M. I has seen image macros! Advice animal memes as visual-verbal jokes[J]. International Journal of Communication, 2016, 10(10): 660-668.

[6] 林爱珺.网络玩梗背后的表达失语与价值观消解[J].人民论坛,2022(4):95-97.

[7] 李欣,彭毅.符号化表演:网络空间丧文化的批判话语建构[J].国际新闻界,2020,42(12):50-67.

[8] 谢卓.从国际冲突“网络玩梗”谈网民媒介素养[J].新闻前哨,2022(12):14-15.

[9] GALIPEAU T. Theimpact of political memes: a longitudinal field experiment[J]. Journal of Information Technology & Politics, 2023, 20(4): 437-453.

[10] MCLOUGHLIN L, SOUTHERN R. By any memes necessary? Small political acts, incidental exposure and memes during the 2017 UK general election[J]. British Journal of Politics & International Relations, 2021, 23(1): 60-84.

[11] LUKÁCS G. Internet memes as protest media in populist hungary[J]. Visual Anthropology Review, 2021, 37(1): 52-76.

[12] TIAN Bian, XIAO Xi, XU Tingyang, et al. Rumor detection on social media with bi-directional graph convolutional networks[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2020: 549-556.

[13] MUNYOLE L A, LI Weimin, LI Shaohua, et al. Dual emotion based fake news detection: a deep attention-weight update approach[J]. Information Processing & Management, 2023, 60(4): 103354.

[14] YAO S. A semantic analysis of the top ten network buzzwords of China in 2020 from the perspective of the prototype theory[J]. Studies in Linguistics and Literature, 2021, 5(4): 61.

[15] PACIELLO M, D'ERRICO F, SALERI G, et al. Online sexist meme and its effects on moral and emotional processes in social media[J]. Computers in Human Behavior, 2021, 116: 106655.

[16] HOFER M, SWAN K. Digital image manipulation: a compelling means to engage students in discussion of point of view and perspective[J]. Contemporary Issues in Technology and Teacher Education, 2005, 5(3/4): 290-299.

[17] 仲兆满,李恒.新媒体环境下突发事件识别与分析研究综述[J].江苏海洋大学学报(自然科学版),2022,31(2):78-88.

[18] LAWLESS T J, O'DEAC J, MILLER S, et al. Is it really just a joke? Gender differences in perceptions of sexist humor[J]. Humor, 2020, 33: 291-315.

[19] APRIANTI R, GINTING E. Sexist humor as a form of sexual violence and prevention effort from an islamic perspective[J]. Psikis: Jurnal Psikologi Islami, 2022, 8(2): 239-250.

[20] MALLYA D R. Comic memes and sexist humor in India: tools for reinforcement of female body-image stereotypes[J]. Rupkatha Journal on Interdisciplinary Studies in Humanities, 2021, 13(4): 1-11.

[21] SCHLOSBERG H. Three dimensions of emotion[J]. Psychological Review, 1954, 61(2): 81.

(责任编辑:李琴 实习编辑:张昌保)