基于多特征融合的在线论坛用户心理健康自动评估

刘德喜 夏先益 万常选 刘喜平 江腾蛟 付 淇

(江西财经大学信息管理学院 南昌 330013)

(江西财经大学数据与知识工程江西省高校重点实验室 南昌 330013)

摘 要 心理健康问题会对社会和谐和家庭幸福造成严重破坏,提前发现有心理健康问题的潜在患者,有利于对其进行及时辅导和治疗.人们利用互联网或社交网络交流沟通、表达情感和观点,这为心理健康的观察提供了新的窗口.本文提出基于多特征融合的在线论坛用户心理健康自动评估框架 F^3TMH,该框架采用贪婪法 F^3TMH_G、投票法F^3TMH_V、后期融合法 F^3TMH_L 和降噪自编码器法 F^3TMH_DA 四种特征融合策略,融合帖子(或其作者)的行为与属性特征、语言或用词风格特征、内容特征(N-Grams 特征、主题特征、词向量特征)、上下文特征,对论坛中帖子所反映的用户(心理康健状况)需要干预的紧急程度(*crisis*:非常紧急,*red*:紧急,*amber*:不紧急,*green*:不需要任何干预)进行自动评估.在 CLPsych2017 shared task 评测任务所提供的数据集上,考察了各类特征、不同的特征融合策略对心理健康自动评估性能的影响.实验发现,相对于行为与属性特征和语言特征,内容特征表现更好,其中基于 Word2Vec 的词向量特征表现最佳,其 *Non-green*(*crisis*、*red*、*amber* 三类)的 *F*1 均值达到 0.429.尽管单独使用行为与属性特征表现不佳,但该特征对 *crisis* 类帖子的识别影响很大,在融合所有特征的基础上去掉该特征后会导致 *crisis* 类帖子的 *F*1 值下降 19.7%.实验还显示,多种类型特征的融合较单一类型的特征表现更优,特征融合后 *Non-green* 的 *F*1 值(0.479)较单一最优特征(0.429)提高 11.6%.各种特征融合策略各有优势,例如,后期融合策略 F^3TMH_L2 更有利于识别心理健康危机程度较高的用户(*crisis* 和 *red* 类帖子),*Urgent* 的 *F*1 值达到 0.608,而 F^3TMH_L 则更有利于识别 *crisis* 类的帖子,自编码融合策略 F^3TMH_DA 对于识别数据量相对较多的 *Flagged* 类(所有非 *green* 类的并集)帖子更有优势,其 *F*1 值达到 0.872.最后还探讨了上下文信息对用户心理危机程度识别的影响.此外,F^3TMH_V 参加了 CLPsych2017 shared task 评测,在官方对参赛系统排名的评价指标 *Non-green F*1 上得分 0.467,排名第一,优于采用深度学习等其他模型和特征的参赛系统.

关键词 在线论坛用户;心理健康自动评估;行为与属性特征;语言特征;内容特征;多特征融合

中图法分类号 TP18 **DOI 号** 10.11897/SP.J.1016.2019.01553

Mental Health Assessment for Online Forum Users Based on Multi-Feature Fusion

LIU De-Xi XIA Xian-Yi WAN Chang-Xuan LIU Xi-Ping JIANG Teng-Jiao FU Qi

(*School of Information Technology, Jiangxi University of Finance and Economics, Nanchang* 330013)

(*Jiangxi Key Laboratory of Data and Knowledge Engineering, Jiangxi University of Finance and Economics, Nanchang* 330013)

Abstract Mental health problems cause serious damage to social harmony and family happiness. Early detection of potential patients with mental health problems is conducive to timely counseling and treatment. Internet or social networks are convenient for people to communicate and express their feelings and opinions, which provides an innovative perspective for mental health detecting. We propose F^3TMH, a framework classifying the posts based on multi-feature fusion, to assess the urgency (*crisis*: very urgent, *red*: urgent, *amber*: not urgent, *green*: no intervention

收稿日期:2017-11-27;在线出版日期:2018-09-18.本课题得到国家自然科学基金(61762042,61363039,61562032)、江西省落地计划项目(KJLD14035)、江西省自然科学基金资助项目(20171BAB202021,20152ACB20003)资助.**刘德喜**,博士,教授,博士生导师,中国计算机学会(CCF)会员,主要研究领域为社会媒体处理、信息检索、自然语言处理.E-mail:dexi.liu@163.com.**夏先益**,硕士研究生,主要研究方向为社会媒体处理.**万常选**,博士,教授,主要研究领域为 Web 数据管理、情感计算.**刘喜平**,博士,副教授,主要研究方向为 Web 数据管理.**江腾蛟**,博士,讲师,主要研究方向为情感计算.**付 淇**,博士研究生,主要研究方向为社会媒体处理.

required) of the mental health status of online forum users. Four kinds of features, the behavior & attribute feature of a post or an author, the linguistics feature, the content features (including N-Grams feature, topic feature and word embedding feature), and the context feature of the target posts, are fused by four strategies in F^{3}TMH, greedy based F^{3}TMH_G, voting based F^{3}TMH_V, late fusion based F^{3}TMH_L and denoising autoencoder based F^{3}TMH_DA. We examine the performance of various features and different feature fusion strategies on the triage dataset provided by the CLPsych2017 shared task. The experimental results show that the content features perform better than the behavior & attribute feature and the linguistics feature. The word embedding feature based on Word2Vec model has the best performance, where the *Non-green* *F*1 score (the mean of *crisis* *F*1, *red* *F*1 and *amber* *F*1) achieved 0.429. Although the behavior & attribute feature of a post or an author does not perform well when use it alone, the *crisis* *F*1 decreased by 19.7% if it is removed from F^{3}TMH_G. Compared with a single feature, multi-feature fusion has a great effect on F^{3}TMH, which *Non-green* *F*1 score (0.479) is 11.6% higher than the best single feature (0.429). Four feature fusion strategies have their own advantages. For instance, F^{3}TMH_L2 performs better to identify the users with higher mental health crisis (*crisis* and *red* posts), which *Urgent* *F*1 reached 0.608, whereas F^{3}TMH_L performs better to identify the *crisis* posts, and F^{3}TMH_DA has advantages in identifying *Flagged* posts (the union of all no *green* posts), where *Flagged* *F*1 is 0.872. Finally, we analysis the effect of the context on mental health assessment for online forum user. Moreover, F^{3}TMH_V participated in the CLPsych2017 shared task and ranked first with the official metrics *Non-green* *F*1 score of 0.467, outperforming other feature engineering based or deep learning based models.

Keywords online forums user; mental health assessment; behavior & attribute feature; linguistics feature; content features; multi-feature fusion

1 引 言

根据世界卫生组织统计,心理健康疾病已成为世界第四大疾病,预计到 2020 年将上升至第二位.全球仅受抑郁症这一心理健康疾病困扰的人口就超过 3 亿.心理健康疾病极大地危害着人们的身心健康[1].心理疾病是导致自杀的主要原因,据报道有超过三分之二的自杀事件是由心理疾病导致的[2-3],给家庭幸福和社会和谐发展带来巨大挑战.及时发现有自我伤害倾向、自杀或抑郁等心理健康问题的个体,有助于尽早提供心理辅导或治疗,具有重要社会意义[4].

传统的心理健康评估,特别是大规模的心理健康评估,多采用基于自评量表的问卷调查方法.这种方法适时性不强,受被试个体填写自评量表当前心理状态影响较大,并且其侵入性的特点也会引起被试个体的抵抗,增加误报率.伴随着互联网技术的逐渐成熟,社交网络平台如 Twitter、Facebook、新浪微博、人人网、微信朋友圈、QQ 空间、在线论坛等,近年来飞速发展,已经成为人们日常生活必不可少的一部分.人们通过社交网络交流信息、抒发情感、发表评论、记录生活中的点点滴滴.社交网络平台也经常被用来表达社交网络用户(以下简称"用户")自己内心的真实情感、心理状态,并且在线寻求帮助,这使得社交网络成为研究抑郁、自我伤害等心理健康问题以及产生原因的一个重要来源.例如,在线论坛 ReachOut① 就是一个围绕心理健康问题进行交流和寻求帮助的网络平台[5-12],该论坛的各个版主由经过专业训练的专家或者具有丰富心理健康咨询经验的志愿者组成,其作用是在线识别存在心理康健问题的用户,并及时给予这些用户相应的帮助和支持.然而,对于有限的专家和志愿者,难以以纯人工的方式及时从每日数以千计的用户帖子中识别出有心理健康问题的用户.

对社交网络用户的研究发现,通过分析用户在社交网络平台上的自述内容以及行为表现有助于及

① https://forums.au.reachout.com/

时发现用户的心理健康状态[13-14]，并能通过一定措施给予干预或辅导. 利用在线论坛等各类社交网络平台上数据自动评估用户心理健康状况，吸引着越来越多来自计算机科学和心理学领域的学者. 尽管已取得了丰富的成果，但心理健康自动评估的准确率还有很大的提升空间，特别是自动评估模型、用于自动评估的特征等方面还有待进一步探索.

本文针对在线论坛用户心理健康自动评估问题，提出基于多特征融合的在线论坛用户心理健康自动评估框架 F^3TMH(Feature Fusion Framework for automatically Triaging Mental Health)，考虑包括行为特征、属性特征、语言特征、内容特征、上下文特征等多种类型的特征，设计四种特征融合策略：贪婪法(Greedy based F^3TMH，F^3TMH_G)、投票法(Voting based F^3TMH，F^3TMH_V)、后期融合法(Late Fusion based F^3TMH，F^3TMH_L)、降噪自编码融合法(Denoising AutoEncoder based F^3TMH，F^3TMH_DA). 以 CLPsych 2017 shared task① 评测发布的 ReachOut 论坛数据集为实验数据，分析了各类特征以及不同的特征融合方法在心理健康评估任务上的效果，并与包括深度学习在内的方法进行对比. 其中基于投票法的特征融合策略 F^3TMH_V 参加了 CLPsych 2017 shared task 评测，从评测结果看，该方法具有较强的竞争力，在 19 家参赛单位 16 支参赛队伍提交的 251 个参赛系统中排名第一.

本文引言部分，介绍心理健康自动评估问题的研究背景；第 2 节相关工作，总结近几年关于社交网络用户心理健康研究现状以及研究的常规方法和相关成果；第 3 节介绍基于多特征融合的在线论坛用户心理健康自动评估框架，主要包括数据预处理、特征工程、多特征融合三个步骤；第 4 节通过丰富的实验考察包括上下文特征在内的不同特征对用户心理健康评估的影响以及不同特征融合策略的实验效果，并与深度学习的实验结果进行对比；最后 1 节对全文进行总结并提出下一步的工作.

2　相关工作

随着互联网和社交网络的普及以及自然语言处理技术日益成熟，基于互联网或社交网络的用户心理健康评估得到了计算机科学领域和心理学领域学者们的广泛关注[2-20]. 相关研究工作中，所采用的方法主要是基于分类的方法. 首先通过用户填写自评量表并采集其社交网络数据，或者雇用心理学专家直接分析社交网络数据，构建心理健康自动评估的训练集，然后提取候选特征，训练分类模型，实现大规模社交网络用户心理健康的自动评估和检测. 在基于分类的心理健康自动评估方法中，特征选择与融合起着决定性的作用. 已有的研究中，用于心理健康自动评估的特征可以概括为四类：行为特征、属性特征、内容特征、社会关系特征.

(1) 行为特征

行为特征刻画了用户在社交网络上的行为表现. 例如，用户在社交网络平台上发帖、点赞、评论、关注等行为，反映在发帖频率、点赞或评论频率(或数量)、在社交网络平台上的活跃时间段、关注(好友)数量等特征上. 多项研究发现[15-18]，个体心理健康状态会影响用户的在线行为. 管理等人[15]在线招募用户填写问卷，根据个体自杀问卷量表得分情况将用户分为高自杀倾向的用户和低自杀倾向的用户，分析他们在微博使用行为上表现的差异性，发现两组用户的社交活跃度、集体关注度、夜间活跃度均不符合正态分布($P<0.05$)，但两组用户行为特征的差异却具有统计学意义($P<0.05$). 为了进一步探讨不同心理状态的用户在社交网络行为上的差异，管理等人[16]还分析了已自杀死亡的用户(自杀死亡组)与无自杀意念的用户(对照组)在微博使用行为上的差异，得出了更具体的结论，即自杀死亡组的微博链接率和微博互动率均低于对照组，自我关注程度则高于对照组.

Bai 等人[19]通过新浪微博用户的网上数据，验证了传统理论中人格和心理健康的相关性，并采用多任务回归学习方法预测新浪微博用户的心理健康状况，研究结果表明，心理健康问题可以通过网络行为表现出来，通过分析用户微博的使用情况，来预测用户的抑郁和焦虑程度是可行的. 在使用社交网络的时间方面，Conner 和 Choudhury 等人[20-22]发现抑郁症用户在夜间更为活跃.

(2) 属性特征

属性特征描述了社交网络用户的个人基本信息，包括年龄、性别、所属地区、用户昵称、职业、用户头像等[13,23-24]，它通常是用户在注册社交网络账户时填写的. Wei 等人[25]在通过社交网络信息研究用户的人格特质时，发现内向型(Introverts)人格在头像中喜欢掩盖自己的脸或者只显示侧面；开放活跃型(Openess)人格喜欢使用和他们朋友一起的合影

① http://clpsych.org/shared-task-2017/

作为头像. Conner 和 Choudhury 等人[20-22]使用用户的年龄、性别等属性特征和标签数、粉丝数等特征计算心理健康状态,发现在抑郁患者中女性用户较男性用户多.

(3)内容特征

语言表达(口语或者是书面语)是人与人之间最主要的交流方式,在一定程度上反映了内心真实的想法,因此,可以将用户在社交网络上发布的帖子内容作为研究用户心理健康的重要依据. 大量研究发现,不同心理健康状态的用户在语言表达、用词风格等方面有着较大差异. Wang 等人[26]通过对中文微博数据集的分析发现,抑郁用户使用第一人称单数较第一人称复数更多,使用负面情感词较正向情感词更多.

使用社交网络文本内容分析用户心理健康时,最常见的方法是对 LIWC(Linguistic Inquiry and Word Count)①词典中的词类进行词频统计[5,10,17,25,27]. 考虑到词对文本的区分度,Brew[7] 等人在 Rey-Villamizar[6]研究的基础上,结合 TF-IDF 计算文本中词的权重,在检测有自我伤害倾向的用户时得到了更高的准确率和召回率. 在基于 LIWC 词典的统计分析方法中,大都是将 LIWC 中所有的词类当作候选特征,没有考量它们对不同心理健康状态的用户的区分能力,而无区分能力的词类(特征)会降低自动评估模型的性能.

基于 LIWC 词典进行词频统计和基于 TF-IDF 的词项权重设置都无法准确地刻画词的同义、近义、歧义等问题,无法准确描述帖子的上下文信息或主题信息. 随着深度学习、主题模型在自然语言处理领域的广泛应用,基于深度学习的词向量模型 Word2Vec 也被心理健康分析的研究者用来表示词特征,从而更好地利用上下文信息[28]. 还有研究者将维基百科等外部知识作为语料库,来训练 LDA (Latent Dirichlet Allocation)主题模型,分析用户帖子的主题[10]. Zhang 等人[29]使用 LIWC 词典和 LDA 主题模型提取特征并训练机器学习算法模型来预测具有自杀意图的用户微博,结果表明,LDA 主题模型能发现与自杀相关的话题,并且能够提高模型的预测性能.

(4)社会关系特征

在社会学中,社会关系是人们在生产和共同生活中形成的人与人之间的相互关系. 在社交网络中社会关系体现在用户关注的好友、关注用户的粉丝、关注的亲密度等. Wang 等人[30]将每个用户看作一个节点,通过节点与节点间的共同朋友、共同标签、交互中的情感倾向等信息来计算节点与节点之间的连接权重,在文献[26]的基础上,加入用户节点间连接权重这一特征,使得抑郁用户检测结果的准确率提高了 15%. Choudhury 等人[21]从社会交流、以自我为中心的社交图来分析,发现抑郁症患者表现出较少社交活动参与,朋友和粉丝较少. Choudhury 等人[31]发现社会隔离的增加(体现为在 Facebook 上减少社交活动和交互)和社会资本(social capital,体现为社会信任、归属感、互助等)的降低能够较好地预测产后抑郁症.

分类问题中的特征融合策略主要有两类,一类是前期融合,即在模型建立前,将样本的各类特征直接拼接或者通过加权拼接,得到样本的综合特征,用于模型的训练;另一类是后期融合,先根据样本的各类(单一)特征训练多个模型,然后将多个模型预测结果进行融合,得到最终的预测结果.

以上内容不能完全概括近几年通过社交网络研究用户心理健康的全部工作. 比如,Liu 等人[27]通过深度学习方法构建模型,用于社会网络用户的人格预测. Yates 等人[32]使用卷积神经网络,对 Reddit 和 ReachOut 论坛用户进行抑郁和自我伤害倾向检测,从而避免了繁琐的特征构建工作. 另外,Manikonda 等人[33]还使用计算机视觉技术来分析用户在社交网络上 Instagram(照片墙)中照片的颜色、内容,以及与照片相关的文本中 LIWC 词类特征和主题特征,发现用户也常在 Instagram 中以分享照片的方式来表达内心的抑郁,并寻求帮助.

2014 年起,计算语言学年会(ACL)专门增加了关于计算语言学与临床医学的专题讨论会 CLPsych (The Computational Linguistics and Clinical Psychology Workshops),旨在把语言学技术应用于心理健康领域,结合计算语言学、自然语言处理技术和临床医学的交叉知识,来探讨利用社交媒体、在线社区论坛等大众平台上的短文本检测用户心理健康的相关理论和方法,吸引了全球各地感兴趣的学者的参与,推动了本领域的发展.

3 心理健康自动评估框架 F³TMH

3.1 问题定义

对于在线论坛帖子集合 $D=\langle P,H,R\rangle$, $P=$

① https://liwc.wpengine.com/

$\{p_1, p_2, \cdots, p_N\}$表示 N 个帖子，含帖子的作者、发布时间等信息，$H=\langle h_1, h_2, \cdots, h_L\rangle$表示 L 个主题块(Thread)，每个主题块中包含若干个帖子，R 是帖子与主题块之间的关系，$r_{i,j}$表示帖子 p_i 属于主题块 h_j. 帖子本身的内容及所在的主题块能在一定程度上反映帖子作者的心理健康状况，需要心理健康专家或论坛版主给予不同紧急程度的关注，该紧急程度共分四个等级，表示为 $C=\{crisis$：非常紧急，red：紧急，$amber$：不紧急，$green$：不需要任何干预$\}$.

本文对于在线论坛用户心理健康评估的定义是：对于 D 中的帖子 p_i，寻找一个映射函数或分类模型 m 和一组特征 F，使得 $m(F(p_i, H, R, O))\in C$ 的结果能正确反映帖子 p_i 需要被心理健康专家或论坛版主关注的紧急程度，也就是帖子作者的心理健康状况. 其中特征 F 可以来自帖子 p_i 本身，也可以来自 p_i 所在的主题块 h，或者其它外部数据资源 O.

关于在线论坛用户心理健康自动评估中多特征融合的问题有学者做了类似的研究. Cohan 等人[8]以帖子的文本内容特征作为 baseline，在此基础上描述了其他特征对实验结果的影响以及增加其他特征后实验结果的变化情况；Friedenberg Meir 等人[12]同样也描述了不同特征组合对实验结果的影响，二者都只是通过简单的特征拼接来探讨多模态特征融合对实验结果的影响，没有系统地探讨多特征线性融合以及非线性融合对心理健康自动评估的影响.

本节提出的基于多特征融合的在线论坛用户心理健康自动评估框架 F³TMH，主要包括数据预处理、特征工程、特征融合三个阶段，如图 1 所示.

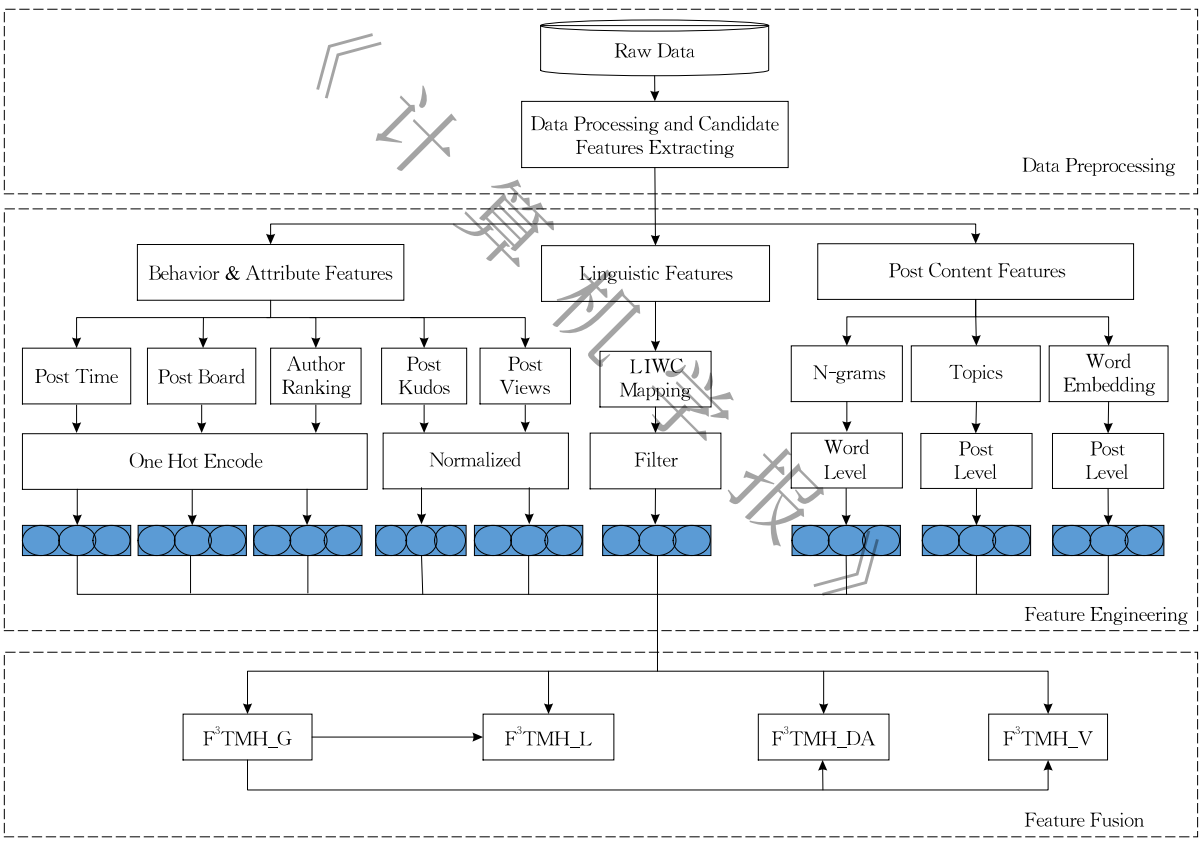


图 1　基于多特征融合的心理健康自动评估框架 F³TMH

其中数据预处理是指统一规范化数据表示，主要包括数据转换、表情符标注和数据过滤等；特征工程是在对数据统计分析的基础上，提取有助于分类的候选特征；特征融合是将提取的候选特征通过贪婪法、投票法、后期融合法、基于深度学习的降噪自编码融合法等进行融合并构建分类模型.

3.2　数据预处理

数据预处理主要包括数据转换、表情符标注和数据过滤三个操作. 数据转换：包括大小写转换和链接的替换，其中后者是将帖子中出现的链接统一替换为“URL”，不对链接对象的类型或内容进行分析. 表情符标注：社交网络的多样性使得表情符成为用户表达内心情感的一种简单直接的方式，统计发现 $crisis$ 类和 red 类样本中多出现负面消极的表情符号，而 $green$ 类中多为积极的表情符，因此将仅出现在训练数据集的 $crisis$ 类或 red 类(合称为 $urgent$ 类)样本中的表情符统一标注为负向，将仅出现在 $green$ 类或 $amber$ 类(合称为 $non\text{-}urgent$ 类)

样本中的表情符统一标注为正向，若表情符既在 *urgent* 类中出现又在 *non-urgent* 类中出现，统一将其标注为中性，如表 1 所示. 数据过滤：主要是去除文本中无法准确解码识别的非英文、非数字、非标点符号如‘□’、‘ぷ’等信息.

表 1 表情符标注示例

类标签	类中的表情符示例	极性
urgent	😕, 😟, 😑, 👀, 😈, 😳, 😟	负向
non-urgent	😂, ❤, 😁, 😊, 😂, ♡, 🐰, 😁, 💗, 👍	正向
urgent & *non-urgent*	😟, 😐, 🙏	中性

不在预处理中进行停用词去除操作. 一方面因为停用词在不同类别的样本中分布不同，能反映出不同心理健康类别的用户的用词习惯，对分类能起到帮助作用；另一方面，停用词的使用也会改变文本的情感倾向性，例如，对数据集的统计发现“not”在 *crisis* 类样本中出现的次数比在 *green* 要多，而“to be or not to be”则直接表明了帖子作者的态度.

3.3 特征工程

F^3TMH 融合了行为特征、属性特征、语言特征、内容特征和上下文特征，其中内容特征又采用主题模型、词向量等多种模型来构建.

3.3.1 行为与属性特征(Behavior & Attribute Features，BAF)

由于数据集中帖子作者的相关信息较少，因此，本文中帖子作者的行为特征仅体现在帖子发布的时间(时段，Post Time)和帖子所在的版块(Post Board)上. 属性特征除了包含相关研究中提到的作者的属性外，还包括帖子自身的属性，具体体现在帖子被点赞的次数(Post Kudos)、帖子被查看次数(Post Views)和帖子作者在 ReachOut 论坛中的等级(Author Ranking).

图 2 显示了不同类别的帖子发帖时间在一天 24 小时中的分布情况. 可以看出 *red* 类和 *crisis* 类的帖子发布在 5:00AM 到 6:00AM 时段的比例，远高于 *amber* 类和 *green* 类帖子发布在该时间的比例，而 *crisis* 类和 *amber* 类的帖子发布在 11:00AM 时段的比例远高于 *green* 类和 *red* 类的帖子发布在该时段的比例.

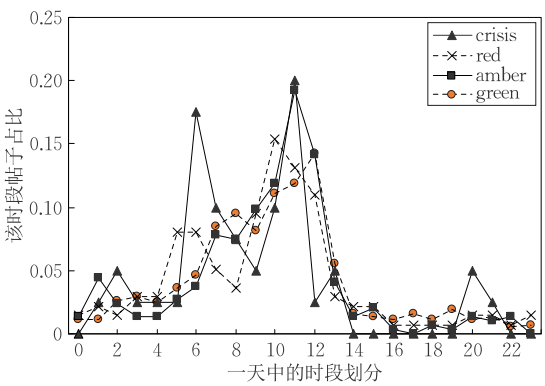


图 2 各类帖子的发布时间分布

图 3 为各类样本被查看次数的分布，其中 *crisis* 类中每条帖子平均被查看 344.5 次，*green* 类中每条帖子平均被查看 273.5 次，说明该论坛上的用户更加关注 *crisis* 类的帖子(心理健康危机程度高的用户).

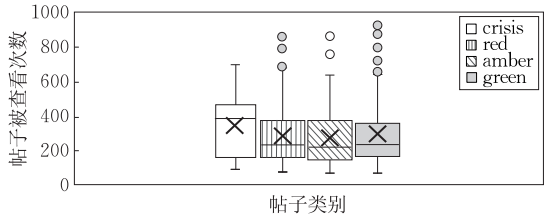


图 3 各类帖子平均被查看的次数分布

图 4 则是从另一角度说明了这一问题，获得一次以上点赞概率最大的是 *green* 类帖子，而 *crisis* 类和 *red* 类帖子不被点赞的概率较大.

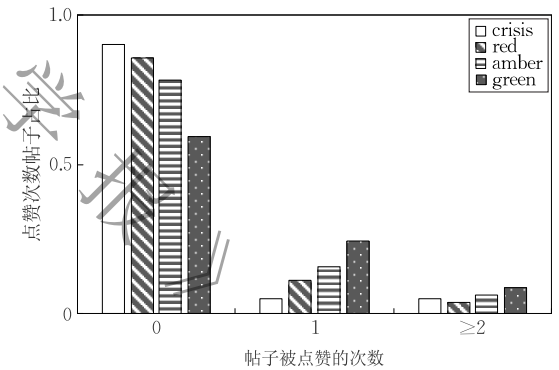


图 4 各类帖子点赞次数的占比分布

经过类似的统计分析后，F^3TMH 中最终选用的行为与属性特征包括：帖子发布时间(时段)、帖子被点赞的次数、帖子被查看的次数、帖子作者在论坛中的等级、帖子所在的版块.

由于用户行为和帖子属性的特征值差异较大，因此采用不同的编码方式. 对于帖子发布时间(时段)、帖子作者在论坛中的等级、帖子所在的板块等标称或是区间类型数据采用 one-hot 编码. 考虑到帖子被点赞的次数和帖子被查看的次数都是数值型特征，而且波动性较大，因此对它们按照式(1)进行 z-score 规范化.

$$x'=(x-\mu)/\sigma \tag{1}$$

其中，x 表示帖子被点赞的次数或者帖子被查看的

次数，μ 和 σ 为相应的均值和标准差，x' 为规范化后的特征值.

3.3.2　语言特征(Linguistic Features，LGF)

大量研究工作[5,19,26]利用 LIWC 词典探索具有不同心理健康状况或不同人格的个体所撰写的日记、博客、微博等文本之间存在的用词差异，并基于这种差异来检测或预测个体的心理健康状况或人格. LIWC 词典包含有 64 个词类，如否定词、生理过程词、社会过程词、认知过程词、人称代词等. 通常的做法是，将每一个词类当作一个特征，文本中属于该词类的词语出现的概率或比例作为相应的特征值. 然而，并不是所有的词类都对心理健康自动评估模型有益，因此，本文在标准差分析的基础上对 64 个词类进行筛选，目的是仅保留对不同类别样本区分能力较强的或者在不同类别样本上差异较大的词类.

基于标准差分析的语言特征筛选步骤如下：

(1) 计算词类的频率. 计算 LIWC 中各个词类在不同心理健康类别样本中出现的频率. 具体地，对于训练集中样本(帖子)$\{p^i=\langle t_1^i,t_2^i,\cdots,t_k^i,\cdots,t_{n_i}^i\rangle, i=1,\cdots,|D|\}$，其中 $|D|$ 表示样本数，n_i 表示帖子 p^i 中的文本长度(词数)，t_k^i 表示帖子 p^i 中的第 k 个单词，LIWC 词典中词类 l 在样本类别 $c\in\{crisis,red,amber,green\}$ 中出现的频率 $TF(l,c)$ 定义如式(2).

$$TF(l,c)=\frac{\sum_{i=1}^{|D|}\sum_{j=1}^{n_i}1_l(t_j^i)\cdot 1_c(p^i)}{\sum_{i=1}^{M}n_i\cdot 1_c(p^i)},$$
$$1_A(x):=\begin{cases}1, & x\in A\\ 0, & x\notin A\end{cases} \tag{2}$$

(2) 计算词类的标准差. 计算 LIWC 中各个词类在不同心理健康类别样本中出现频率的标准差，用来衡量 LIWC 词类在不同心理健康类别上的差别. 标准差越大，说明该 LIWC 词类越能有效地区分样本所属的心理健康类别. 由于不同的 LIWC 词类在数据集上的频率绝对差异较大，因此，在标准差计算时，利用该 LIWC 词类在不同心理健康类别中频率的最大值进行归一化. 具体计算如式(3)所示.

$$\delta_l^2=\sum_c(TF'(l,c)-\mu)^2,$$
$$TF'(l,c)=\frac{TF(l,c)}{\max_c(TF(l,c))},$$
$$u=\frac{\sum_c TF'(l,c)}{\sum_c 1} \tag{3}$$

其中，$TF'(l,c)$ 表示对 $TF(l,c)$ 进行归一化，μ 表示词类 l 在所有四类样本上归一化频率的平均值.

(3) 词类筛选. 选择 δ_l 最高的 top-k 个 LIWC 词类作为语言特征. 实验按照 δ_l 排序依次选择前 k 类词作为候选特征，发现 $k=28$ 时的效果最好，因此在 F^{3}TMH 中，仅从 64 个 LIWC 词类中选择 28 个最有区分度的词类作为语言特征. 表 2 是 δ_l 排名前 10 的 LIWC 词类，包括人称代词(如：You_c、We_c)、生理过程词(如：Sexual_c、Ingest_c、Anger_c)、社会过程词(如：Relig_c、Death_c)、认知过程词(Money_c、Friends_c)等，这与文献[29,34]得出的结论基本一致. 最后，训练样本或测试样本 p^i 的语言学特征中，对应词类 l 的特征值为 $LGF(l,p^i)$，如式(4)所示.

$$LGF(l,p^i)=\frac{1}{n_i}\cdot\sum_{j=1}^{n_i}1_l(t_j^i) \tag{4}$$

表 2　部分 LIWC 词类在四类样本上的标准差

排名	词类	标准差	排名	词类	标准差
1	Relig_c	0.4330	6	Ingest_c	0.3247
2	Death_c	0.3861	7	Assent_c	0.3192
3	You_c	0.3787	8	Sexual_c	0.2846
4	Friends_c	0.3658	9	Money_c	0.2760
5	We_c	0.3518	10	Anger_c	0.2469

3.3.3　内容特征(Post Content Features，PCF)

不同心理健康状况的个体所关注的主题和内容会有较大的差异，例如，有自杀倾向的用户较少地表达与未来相关的内容，而是更多地表达与死亡相关的主题[35]. F^{3}TMH 中，内容特征通过 N-Grams 特征、主题特征和词向量特征来表达.

(1) N-Grams 特征(N-Grams Features，NGF). 该特征是从词的粒度上来分析短文本的内容，将文本按照 N-Grams 切分为词或者词组，并计算其 TF-IDF 权重作为特征值. 考虑到 uni-gram 无法准确表示用户表达的情绪或其它语义，如“not pleasure”、“have fun”等，本文中同时使用 uni-gram 和 bi-gram 两个粒度的特征.

(2) 主题分布特征(Topic Features，TPF). 由于一词多义或者近义等问题，从词的粒度来分析帖子内容，会存在偏差. 主题特征能够很好地解决该问题，在降维的同时，提取出关键的主题表示帖子内容. 本文使用 CLPsych2017 shared task 提供的全部数据训练 LDA 主题模型. 在主题建模之前，去除文档(帖子)中出现次数少于 5 的词，然后采用 Python 的第三方库 gensim① 来实现 LDA 主题分析. 将训

① http://radimrehurek.com/gensim/

练集和测试集中每个帖子中的主题分布概率作为主题特征，记为 $TPF(p^i)$，如式(5)所示.

$$TPF(p^i)=\langle T_1^i,T_2^i,\cdots,T_j^i,\cdots,T_k^i\rangle \quad (5)$$

式中，T_j^i 表示帖子 p^i 中第 j 个主题的概率. 对于主题个数 k 的选取，在尝试 $k=20,25,30,80$ 后发现，对于本文采用的数据集，主题个数 k 为 25 时效果最好.

(3) 词向量特征(Word Embedding Features, WEF). 基于神经网络和深度学习的词向量模型已经在自然语言处理领域得到了广泛的应用. 本文使用 CLPsych2017 shared task 的全部数据，在去掉文档(帖子)中出现次数少于 5 的词后[36]，利用 python 第三方开源包 gensim 训练 Word2Vec 模型，词向量的维度设置为 300. 得到数据集中每个词的词向量表示后，对于帖子 p^i 中全部词的词向量求平均，得到帖子 p^i 的词向量表示，如式(6)所示.

$$WEF(p^i)=\frac{1}{n}\sum_{j=1}^{n_i}\boldsymbol{v}(t_j^i) \quad (6)$$

式中，$\boldsymbol{v}(t_j^i)$ 表示帖子 p^i 中第 j 个词 t_j^i 的词向量.

3.4 特征融合

考虑到不同特征之间可能存在着互补信息，融合这些特征会提高心理健康自动评估模型的性能. 本文设计了四种特征融合策略：贪婪法(Greedy based F^{3}TMH，F^{3}TMH_G)、投票法(Voting based F^{3}TMH，F^{3}TMH_V)、后期融合法(Late Fusion based F^{3}TMH，F^{3}TMH_L)、降噪自编码器融合法(Denoising AutoEncoder based F^{3}TMH，F^{3}TMH_DA).

(1) 贪婪法特征融合策略 F^{3}TMH_G

贪婪法特征融合策略 F^{3}TMH_G 的基本思想是，首先从五种候选特征{行为与属性特征 BAF、语言特征 LGF、N-Grams 特征 NGF、主题特征 TPF、词向量特征 WEF}中选择一种使得心理健康自动评估(分类)模型效果最好的特征. 在此特征的基础上，从剩下的四种特征中选择一种补充到已经选择的特征集合中，补充的特征一方面要使得补充后模型的性能得到提高，同时还要求该特征较其它三种特征对模型性能改善的效果更好. 重复上述过程，直到没有新的特征加入或者所有特征全部加入特征集合中. F^{3}TMH_G 的算法如算法 1 所示.

算法 1. 贪婪法特征融合策略 F^{3}TMH_G.

输入：训练数据集 D，候选特征集 $\psi=${行为与属性特征 BAF，语言特征 LGF，N-grams 特征 NGF，主题特征 TPF，词向量特征 WEF}

输出：心理健康评估模型 m

1. $G=\varnothing$；
2. WHILE $\psi\neq\varnothing$：
3. $f=\arg\max_{f'\in\psi}F_1(m(G\cup\{f'\}))$；//$m(G)$表示以 G 为特征的模型，$F_1(m)$表示模型 m 的性能
4. IF $F_1(m(G\cup\{f\}))>F_1(m(G))$：//如果融合特征 f 能提高模型的性能
5. $G=G\cup\{f\}$，$\psi=\psi-\{f\}$；
6. ELSE BREAK；
7. RETURN $m(G)$；

(2) 投票法特征融合策略 F^{3}TMH_V

鉴于特征之间的差异性，基于不同特征或特征组合的心理健康评估模型对样本的辨识能力也不相同. 投票法特征融合策略利用多个基于不同特征或特征组合的心理健康评估模型对样本的分类结果，采用投票的方式确定最终的标签. 参与投票的模型来自贪婪法特征融合策略得到中间结果. 投票法特征融合策略要求参与投票的模型性能不能太差，因此本文将参与投票的模型的性能阈值设置为采用单一特征时最优模型的性能，即 $\varepsilon=\max_{f'\in\psi}F_1(m(f'))$. 投票法特征融合策略 F^{3}TMH_V 中，参与投票的模型集合来自算法 2.

算法 2. 参与 F^{3}TMH_V 投票的模型产生算法.

输入：训练数据集 D，候选特征集 $\psi=${行为属性特征 BAF，语言特征 LGF，N-grams 特征 NGF，主题特征 TPF，词向量特征 WEF}

输出：参与 F^{3}TMH_V 投票的模型

1. $f=\arg\max_{f'\in\psi}F_1(m(f'))$，$\varepsilon=\max_{f'\in\psi}F_1(m(f'))$；寻找使得模型性能最优的单一特征及参与投票的模型性能的阈值.
2. $G=\{f\}$，$M=\{m(f)\}$，$\psi=\psi-\{f\}$；
3. WHILE $\psi\neq\varnothing$：
4. FOR f IN ψ：
5. IF $F_1(m(G\cup\{f\}))\geq\varepsilon$：
6. $M=M\cup\{m(G\cup\{f\})\}$；如果融合特征 f 后模型 m 的性能高于阈值，将模型 m 添加到模型集合 M 中
7. $f=\arg\max_{f'\in\psi}F_1(m(G\cup\{f'\}))$
8. $G=G\cup\{f\}$，$\psi=\psi-\{f\}$
9. RETURN M；

对于给定的帖子 p，用模型集合 M 中的各个模型分别对 p 进行分类(贴类标签)，被贴上最多的类标签作为帖子 p 的最终标签. 如果被贴上的两种类标签数量相同，则选择先验概率较大的类标签，即样本较多的类标签，作为 p 的最终标签.

为了考察 F^{3}TMH_V 中分类模型性能的阈值设置(算法 2 中步骤 5)及基于贪婪法特征扩展(算法 2 中步骤 6)的效果，作为对比，本文构建了

F^3TMH_V2，其中参与投票的每个模型只采用单一特征，且不考虑每个模型的实际性能，即 $M_2=\{m(\text{BAF}),m(\text{LGF}),m(\text{NGF}),m(\text{TPF}),m(\text{WEF})\}$.

(3) 后期融合法特征融合策略 F^3TMH_L

投票法特征融合策略 F^3TMH_V 视参与投票的各个模型同等重要，不考虑各个模型之间的差异. 与 F^3TMH_V 不同的是，后期融合法特征融合策略 F^3TMH_L 不直接采用投票的方式为帖子贴上最终的标签，而是训练一个分类器，将多个模型对帖子的分类结果作为该分类器的输入，由分类器代替投票器，确定帖子的最终类标签. F^3TMH_L 如图 5 所示，其中待融合的模型来自算法 2.

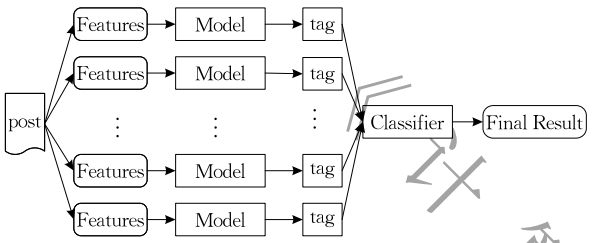


图 5　后期融合法特征融合策略 F^3TMH_L

如果将后期的分类器简化为投票过程的话，从本质上看，F^3TMH_V 是 F^3TMH_L 的一个特例. 类似地，设计后期融合法特征融合策略 F^3TMH_L2，待融合的模型来自 M_2.

(4) 降噪自编码器融合策略 F^3TMH_DA

该特征融合策略是对线性融合后的特征采用深度神经网络无监督地逐层提取原特征的高阶特征. 为了能更好地区分来自其它输入配置的测试样例，迫使自编码器隐层发现更好的鲁棒性特征，选择部分输入特征随机增加噪声来训练自编码器，然后重组高阶特征来拟合原始输入，以期获得原始特征更好的抽象表达形式.

将 3.3 节特征工程中提取的候选特征 $F=\{$行为与属性特征、语言特征、N-Grams 特征、主题特征、词向量特征$\}$通过拼接得到融合后的特征 F'，然后采用去噪自编码器的特征融合方式提取深层次的高阶特征，其结构如图 6 所示.

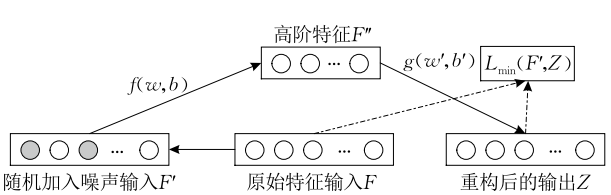


图 6　自编码器特征融合策略

关于特征融合过程中模型 m 的选择问题，本文尝试了 Python 工具库 sklearn 中的多种机器学习方法，包括 Logic Regression、KNN、XGboost、SVM 等，其中 SVM 效果最好，所以实验中的模型 m 主要采用 SVM 分类器. 另外，后期融合法特征融合策略 F^3TMH_L 中用于最终类标签判断的分类器也同样采用 SVM. 在采用降噪自编码器 F^3TMH_DA 融合了基础特征后还使用了深度学习 softmax 分类器和随机森林(RandomForest，RF)分类器.

样本数据分布的不均衡会使得 SVM 的分类结果向着样本数较大的类别倾斜，实验发现其倾斜程度跟样本所占比例呈现一定的正相关关系. 为应对样本不均衡问题，在训练 SVM 时，对每类样本的权重按式(7)进行设置，给样本较少的类别中的样本更大的权重.

$$w_c=\frac{1}{Z}\cdot\frac{|D|}{|D|_c},\ Z=\sum_c\frac{|D|}{|D|_c},\quad c\in\{crisis,red,amber,green\}\tag{7}$$

式中 $|D|$ 为训练数据集中样本数量，$|D|_c$ 为类标签为 c 的样本数量.

SVM 分类器中惩罚因子 C 的大小表示对于离群点的重视程度，核函数 $kernel$ 是对特征空间的映射，C 和 $kernel$ 的选择直接影响样本分类器的性能. 在通过参数优化调优后，将 C 和 $kernel$ 分别设置为 1000 和“rbf”.

4　实验分析

4.1　数据描述

实验数据采用 CLPsych2017 workshop 上的 shared task 提供的 ReachOut 在线论坛帖子，帖子包含了用户发帖时间、发帖作者名称、帖子所在论坛版块、帖子被查看次数、帖子被点赞数、帖子的文本内容等信息. 数据集包括 2012 年 1 月到 2017 年 3 月期间发表的帖子，它们被分为训练集和测试集. 训练集共有 65 755 个帖子，其中人工标注的帖子 1188 个，未标注的帖子 64 567 个；测试集共有 92 206 个帖子，其中用于评测的人工标注帖子 400 个，另有 91 806 个帖子是与人工标注的 400 个帖子相关(作者相同或者具有回复关系等)的其它帖子.

人工标注示例及标注说明简要介绍如表 3 所示，各类别样本分布情况如表 4 所示. 从表 4 可以看出，各类别样本分布极其不均衡，有心理健康问题的人是极少数，这与实际相符，却也是本研究领域面临的一大挑战.

表 3 帖子标注示例

帖子	标签(triage)	标注说明
I can't do anything right. Why am I even still alive.	*crisis*	The author or someone else is at risk of harm.
I don't think I'm fully over that friendship yet.	*red*	The post should be responded to as soon as possible.
So many things to stay here for, so many things I need to leave behind	*amber*	The post should be responded to at some point, if the community does not rally strongly around it.
I am proud of myself for surviving three weeks surrounded by the same people.	*green*	The post can be safely ignored or left for the community to address.

表 4 样本类别分布

类标签	训练数据	测试数据
Crisis	40(3.4%)	42(10.5%)
Red	137(11.5%)	48(12.0%)
Amber	296(24.9%)	94(23.5%)
Green	715(60.2%)	216(54.0%)
总数	1188	400
没标签数据	65755	91806

4.2 评测指标

评测指标采用 CLPsych2017 shared task 官方评测指标,包括:

(1) *Non-green F*1:*crisis*、*red*、*amber* 三类样本分类结果 *F*1 值的宏观平均.该指标反映模型对非健康类样本的识别能力,以及对其心理健康危机程度的判断能力.该指标是 CLPsych2017 对各参赛系统进行排名的指标.

(2) *Flagged F*1:将所有非 *green* 类的样本视为一个大类,称为 *flagged* 类,*green* 类和 *flagged* 类样本分类结果 *F*1 值的宏观平均.该指标反映模型对健康类样本和非健康类样本的区分能力.

(3) *Urgent F*1:将 *red* 类和 *crisis* 类的样本视为 *urgent* 类,*green* 类和 *amber* 类的样本视为非 *urgent* 类,*urgent* 类和非 *urgent* 类样本分类结果 *F*1 值的宏观平均.该指标反映模型对需要紧急处理和不需要紧急处理的样本的区分能力.

(4) *All F*1:*crisis*、*red*、*amber*、*green* 四类样本分类结果 *F*1 值的宏观平均.

为了与文献[32]进行对比,本文也同时给出 *Flagged*、*Urgent*、*All* 三种评测定义下的正确率(*Accuracy*).

4.3 对单类特征的评测

表 5 是对采用单类特征时的评测效果.总体上看,内容特征(NGF、TPF、WEF)较行为与属性特征(BAF)和语言特征(LGF)在各项评测指标上表现要好,其中用户行为与属性特征(BAF)对用户的心理健康评估表现最差,*Non-green F*1 和 *All F*1 值分别为 0.234 和 0.352,而语言特征(LGF)的 *Non-green F*1 和 *All F*1 值也只有 0.279 和 0.363.内容特征中 NGF 和 TPF 特征表现相当,而 Word2Vec 词向量特征 WEF 表现最好,*Non-green F*1 和 *All F*1 值达到 0.429 和 0.534,这说明 Word2Vec 词向量特征在短文本的用户心理健康评估问题上效果更好.另外,尽管 NGF 和 TPF 特征在 *Non-green F*1 和 *All F*1 值上都不如 WEF 特征,但 NGF 特征在 *Urgent* 正确率(*Acc*)上、TPF 在 *Flagged F*1 上优于其它特征在这些指标上的表现.

表 5 单类特征在用户心理健康评估模型上的评测结果

Feature	*Non-green*	*Flagged*		*Urgent*		*All*	
	*F*1	*F*1	*Acc*	*F*1	*Acc*	*F*1	*Acc*
BAF(行为与属性特征)	0.234	0.735	0.634	0.431	0.400	0.352	0.463
LGF(语言特征,$k=28$)	0.279	0.742	0.601	0.508	0.411	0.363	0.453
NGF(N-gram 特征)	0.352	0.764	0.824	0.455	0.714	0.470	0.657
TPF(主题特征,$k=25$)	0.374	0.819	0.753	0.623	0.675	0.485	0.625
WEF(词向量特征)	**0.429**	0.814	**0.831**	0.584	0.703	**0.534**	**0.670**

对表 5 中单类特征评测效果的分析如下:

(1)本文采用的行为与属性特征 BAF 主要是帖子发布时间、帖子被点赞的次数、帖子被查看的次数、帖子作者在论坛中的等级、帖子所在的版块等,这些特征大都来自单个帖子自身属性,一方面无法反映帖子内容,另一方面也无法准确刻画帖子作者的行为模式.可以预期的是,如果把用户发布的多个或全部帖子统一考虑,刻画出的用户行为模式应该更准确.

(2) N-Grams 特征 NGF 是以 uni-gram(词)和 bi-gram 在帖子中出现的频率为基础,计算 N-Grams 在帖子中的 TF-IDF 权重得到的.由于能够有效地

反映帖子的内容，NGF 特征取得了较好的效果. 然而，来自 ReachOut 论坛的帖子受平台限制，多数帖子内容较短，大部分 N-Grams 在文本中出现次数较少，导致通过 N-Grams 的 TF-IDF 权重区分帖子的类别不够理想.

(3) 语言特征 LGF 以 LIWC 词类在帖子中出现的频率为特征，与 N-Grams 特征存在类似的问题，特别是当把这种频率统计局限在用户发布的长度很短的单个帖子上时，更是无法有效地区别用户(或帖子)的用词特点，导致在用户层面心理健康分析上表现较好的 LIWC 词类统计特征，在帖子层面上却表现欠佳. 可能的改进策略是，用帖子作者发布的全部帖子上的 LIWC 统计特征，对帖子中的 LIWC 统计特征进行修正.

(4) 与 N-Grams 特征 NGF 仅关注帖子自身内容不同的是，主题特征 TPF 和词向量特征 WEF 在整个数据集上分析词之间的关系及帖子内容之间的关系，克服了帖子内容短给分类带来的障碍，取得了良好的效果. 由于涉及用户心理健康的主题丰富多样，训练数据可能无法在主题层面较好地覆盖测试数据，因此，主题特征用于心理健康评估还有较大的提升空间. 词向量特征 WEF 效果相对较好，原因是该特征能很好地利用整个数据集的内容，并且词嵌入的平均表示形式更适用于短文本.

为了更进一步考察每一类特征在用户心理健康评估模型中的作用，本文在融合 BAF、LGF、NGF、TPF、WEF 的基础上，观察去掉某一类特征对评估模型的影响，如图 7 所示，其中 ALLF＝(BAF＋LGF＋NGF＋TPF＋WEF)融合了五类特征. 为了能更清晰地观察到各类特征对 *crisis*、*red*、*amber*、*green* 各类样本的影响，图 7 中展示了评估模型在四类样本上的评测结果，即 *F*1 值，标记为 *crisis_F*、*red_F*、*amber_F*、*green_F*，*F*1-*macro* 则是四类样本评测结果 *F*1 值的宏观平均.

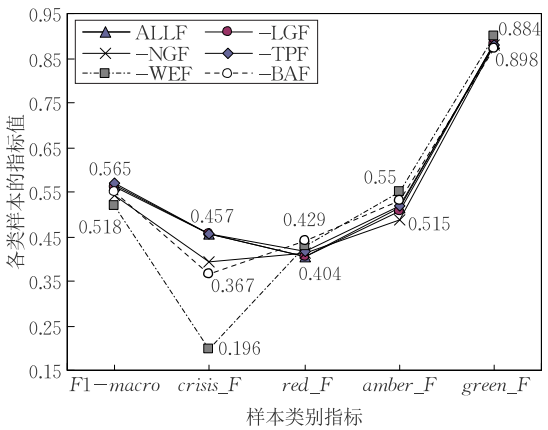


图 7　各类特征对心理健康评估的影响

图 7 中，去掉某类特征后，最大的变化体现在 *crisis* 类样本和 WEF、BAF 和 NGF 三类特征上. 去掉词向量特征 WEF 后，*crisis_F* 从 0.457 下降到 0.196，反映出 WEF 在识别 *crisis* 类样本上的重要作用. 造成这种结果的主要原因在于 *crisis* 类样本的数量过少，通过其他特征较难在太少的样本上归纳或识别出有价值的模式.

对比图 7 和表 5，一个比较有趣的现象是，尽管单独使用用户行为与属性特征 BAF 时效果不如其它四类特征，但 BAF 特征对 *crisis* 类样本却非常重要，去掉 BAF 后，*crisis_F* 从 0.457 下降到 0.367，下降 19.7%. 这说明，尽管 BAF 特征无法有效地刻画全部样本的心理健康状况，但对心理危机程度最高的 *crisis* 类样本却有很好的区分性，原因是此类用户的行为与属性模式与其他用户差异很大，在单个帖子上都展示地较为充分.

图 7 显示，在 *crisis* 类样本上，N-Grams 特征 NGF 较语言特征 LGF 和主题特征 TPF 更不可被替代. LGF 与 NGF 的原理基本一致，但由于使用的词类仅限于 LIWC 中的词，不如 N-Grams 丰富，因此 LGF 基本可以被 NGF 覆盖；主题特征 TPF 和词向量特征 WEF 都是基于整个数据集训练得到，主题特征是从文档的粒度来处理的，而词向量的特征则是从词粒度来处理的，相对来说，词向量特征 WEF 更为细致，基本能够覆盖主题特征 TPF.

图 7 还显示，对 *crisis* 类样本影响较大的 WEF、BAF 和 NGF 三类特征，在 *red*、*amber*、*green* 样本上的表现却不尽相同. 例如，仅从 *amber_F* 和 *green_F* 上看，去掉 WEF 后，*amber* 和 *green* 样本上的评测结果尽管变化不大，但却比去掉 WEF 前稍好. 不过，我们应该小心应对这两类样本上的评测结果，原因是样本严重不均衡可能会带来错觉. 设想，如果去掉 WEF 后，全部样本被分类到 *green* 类(在测试集中占 54%)中，可以使得 *green* 类样本上的召回率和精确率达到 1 和 0.54，相应地 *F*1 高达 0.7，但这没有意义. 这也是 CLPsych2017 官方对参赛系统排名采用 *crisis*、*red*、*amber* 三类样本上 *F*1 值的宏观平均的原因.

从全体样本上看，去掉 WEF、BAF 和 NGF 后，*F*1-*macro* 的下降程度较大，而这三类特征都有较好的代表性，其中，BAF 代表用户行为与属性、NGF 代表帖子表层内容、WEF 代表帖子深层内容.

虽然词向量特征有较好的实验效果，但是对于较长帖子来说，词向量求平均的做法会削弱文本中关键词的作用. 例如，有 *crisis* 类样本中存在一些较为消极的关键字“killing”、“pain”、“self-harm”、“tired”等，但是使用词向量特征的构建模型将其识别为 *amber* 类，原因是该帖子长度为 2769(词数)，平均词向量的表示形式大大的稀释了这些词的作用，导致分类错误. 而利用贪婪法融合多种类型的特征后，能较好地识别这类样本的类别.

4.4 对特征融合策略的评测

表 6 是对四种融合策略的评测结果，各类策略中，用于考察模型 m 性能的指标 $F1(m)$ 均为 *Non-green* $F1$ 值，即 *crisis*、*red*、*amber* 三类样本上 $F1$ 值的宏观平均.

贪婪法特征融合策略 F³TMH_G 中，首先选择表 5 中 *Non-green* $F1$ 值最大的特征 WEF，再根据算法 1 增加特征，依次得到“WEF＋BAF”、“WEF＋BAF＋NGF”，最终，F³TMH_G 选择的特征融合结果是“WEF＋BAF＋NGF＋LGF”.

表 6 显示，采用贪婪法特征融合策略 F³TMH_G 建立的心理健康自动评估模型，在融合了多类特征后，评测结果的大部分指标都有明显的提升，其中融合了 WEF、BAF、NGF、LGF 四类特征时得到最优结果，*Non-green* $F1$、*Flagged* $F1$ 和 *All* $F1$ 分别达到 0.467、0.852 和 0.570，较单个最优特征分别提高了 8.9%、4.7%和 6.7%.

投票法特征融合策略 F³TMH_V，是将 F³TMH_G 特征融合过程中产生的、性能超过某一阈值的模型(本文设定为采用 WEF 特征时的 *Non-green* $F1$ 值 0.429)保留下来，参与最终类标签的投票. 表 6 显示，从采用单个特征 WEF 的模型到采用融合特征“WEF＋BAF＋NGF＋TPF＋LGF”的模型，共有 11 个模型参与最后的投票. 从评测结果看，投票法特征融合策略 F³TMH_V 相对于贪婪法特征融合策略 F³TMH_G 并没有明显的改善，*Non-green* $F1$ 值没有发生变化. 然而，对比基于单类特征构建的分类模型进行投票的策略 F³TMH_V2，F³TMH_V 显然更优，这一方面说明使用投票法特征融合策略时，参与投票的模型自身性能不能太差，另一方面说明贪婪法特征融合策略产生的多个模型尽管较好，但各个模型有趋同性，使得投票策略性能得不到充分体现.

表 6 对融合策略的评测结果

	融合策略	*Non-green*	*Flagged*		*Urgent*		*All*	
		F1	*F1*	*Acc*	*F1*	*Acc*	*F1*	*Acc*
F³TMH_G	**WEF**	**0.429**	**0.814**	**0.831**	**0.584**	**0.703**	**0.534**	**0.670**
	WEF＋BAF	**0.453**	0.856	0.871	0.550	0.629	0.560	0.698
	WEF＋NGF	0.442	0.838	0.880	0.554	0.707	0.550	0.698
	WEF＋TPF	0.435	0.825	0.846	0.560	0.700	0.540	0.680
	WEF＋LGF	0.428	0.814	0.831	0.581	0.692	0.533	0.670
	WEF＋BAF＋NGF	**0.464**	0.849	0.874	0.571	0.648	0.568	0.700
	WEF＋BAF＋LGF	0.455	0.860	0.872	0.550	0.629	0.562	0.700
	WEF＋BAF＋TPF	0.436	0.854	0.866	0.522	0.612	0.547	0.688
	WEF＋BAF＋NGF＋LGF	**0.467**	0.852	0.874	0.571	0.648	0.570	0.703
	WEF＋BAF＋NGF＋TPF	0.456	0.856	0.875	0.553	0.638	0.562	0.698
	WEF＋BAF＋NGF＋TPF＋LGF	0.459	0.859	0.876	0.553	0.638	0.565	0.700
F³TMH_V	F³TMH_V2	**0.394**	0.848	0.860	0.528	0.704	0.514	0.693
	F³TMH_V	**0.467**	0.845	0.877	**0.591**	**0.746**	0.569	0.708
F³TMH_L	F³TMH_L2	0.464	0.852	0.845	**0.608**	0.776	0.566	0.708
	F³TMH_L	**0.468**	0.838	**0.895**	0.545	0.736	0.570	0.710
F³TMH_DA	SVM	**0.469**	**0.871**	**0.878**	0.566	0.652	**0.574**	0.710
	RandomForest	0.258	0.730	0.625	0.293	0.198	0.399	0.635
	Softmax	**0.167**	0.870	**0.955**	0.000	0.000	0.351	0.677

尽管投票法特征融合策略 F³TMH_V 在大部分指标上与贪婪法特征融合策略 F³TMH_G 相当，但表 6 也显示，F³TMH_V 在 *Urgent* $F1$ 和 *Urgent Accuracy* 指标上得分 0.591 和 0.746，较 F³TMH_G 的 0.571 和 0.648 高出 3.5%和 15.1%，说明投票法特征融合策略 F³TMH_V 在识别心理健康危机程度较高的 *crisis* 类和 *red* 类帖子时比较有效.

与投票法特征融合策略 F³TMH_V 不同的是，后期特征融合策略 F³TMH_L 不是简单地对 F³TMH_V 中的 11 个模型结果进行投票，而是基于这些结果训练一个分类器，从而获得更为精细的融合模式. 表 6 显示，总体上看，后期特征融合策略

F³TMH_L 在 *Non-green* *F*1(0.468)和 *All* *F*1 值(0.570)较 F³TMH_V 并没有比较明显的提高(0.467 和 0.569).然而,更细致的观察发现,F³TMH_L 的 *Flagged* *F*1 和 *Urgent* *F*1 分别为 0.838 和 0.545,显著低于 F³TMH_V 在这两个指标上的得分 0.845 和 0.591.考虑到 *Flagged* 是将 *crisis*、*red*、*amber* 三类样本视为一个大类,而 *Urgent* 是将 *red* 类和 *crisis* 二类样本视为一个大类,*Flagged* *F*1 和 *Urgent* *F*1 较低,而 *Non-green* *F*1 值(*crisis*、*red*、*amber* 三类样本上 *F*1 值的宏观平均)反而较高,只能是 F³TMH_L 在 *crisis* 类样本上较 F³TMH_V 有更高的 *F*1 值引起的,说明 F³TMH_L 较 F³TMH_V 在 *crisis* 类样本的识别效果上表现更优.

表 6 还显示,在 F³TMH_L2 中,参与融合的模型是五个基于单类特征构建的,其中如 BAF 等部分特征构建的模型性能效果较差,但使用后期融合策略后,F³TMH_L2 得到的评测结果的总体表现与 F³TMH_L、F³TMH_V、F³TMH_G 基本相当,而 F³TMH_L2 的 *Urgent* *F*1 值达到 0.608,是所有参与比较的特征融合策略在该指标上的最高值.对比 F³TMH_L2 和 F³TMH_V2 发现,如果直接采用比较简单的投票融合策略,当参与投票的模型性能较差时,投票的结果反而不如表现较好的单个模型,而后期融合的策略则能学习到比简单投票更复杂更有效的融合模式.

表 6 中降噪自编码融合策略 F³TMH_DA 使用由 F³TMH_G 选择出的最优特征组合,然后采用降噪自编码提取其高阶特征,再分别采用 SVM、随机森林和 softmax 分类器进行分类.实验结果显示,SVM 分类效果明显优于另两类分类器,其中 SVM 在 *Flagged* 的 *F*1 和 *Flagged* 的 *Accuracy* 指标上得分分别为 0.871 和 0.878,与贪婪法融合 F³TMH_G 的最优得分 0.852 和 0.874 相比提高了 2.2%和 0.4%,*Non-green* 得分 0.469 也高于其他融合策略,说明 F³TMH_DA 法在识别 *green* 和 *amber* 类的帖子上更有优势.基于深度学习的自编码器特征融合方法能更好的提取出数据抽象特征表示,可以用低维特征更好的表示原特征,从而在一定程度上能减少特征维度灾难给分类性能带来的影响.

另外,实验显示,利用降噪自编码提取特征后再用 softmax 和 RandomForest 分类器的效果较差,其中 softmax 虽然在识别 Flagged 类帖子时准确率高达 0.955,但是对于 *Urgent* 类帖子识别为 0,这是因为基于神经网络的 softmax 分类器较为依赖更多的数据信息,而 *Urgent* 类数据较少,同时也说明 SVM 在小规模数据集上有优势.

4.5　上下文信息对心理健康评估的影响

由于帖子文本内容较短,各类特征不够准确丰富,可能导致评估模型失败,本节探讨上下文信息对心理健康自动评估的影响.采用的上下文信息主要为原帖所在主题块(Thread)中的其他帖子.在同一主题块(Thread)中的帖子一般是对某一主题(Subject)或者与之相关的主题的讨论,它们之间具有一定的相关性.本文假设,与原帖相似性高的帖子更有助于原帖类别的判断,因此按照式(8)计算原帖 p 与 p 的上下文 h_p 中候选帖子 $c_i \in h_p$ 之间的相似度,并按照式(9)进行 softmax 归一化.

$$S_{c_i,p}=\frac{\boldsymbol{f}_{c_i}\cdot\boldsymbol{f}_p}{\|\boldsymbol{f}_{c_i}\|\|\boldsymbol{f}_p\|} \tag{8}$$

$$\alpha_{c_i,p}=\exp(S_{c_i,p})\Big/\sum_{c_j\in h_p}\exp(S_{c_j,p}) \tag{9}$$

其中,$\boldsymbol{f}_{c_i}$ 和 $\boldsymbol{f}_p$ 分别表示 c_i 和 p 的特征向量,包括:行为属性特征、语言特征、内容特征等.原帖 p 的上下文特征(Context Feature,CTXF)是所有候选帖按其相似度的加权,如式(10)所示.

$$CTXF=\sum_{c_i\in h_p}\alpha_{c_i,p}\boldsymbol{f}_{c_i} \tag{10}$$

考虑到心理健康自动评估的适时性,上下文信息只选用主题块中原帖 p 发布之前的帖子.

表 7 是增加上下文信息后的实验评测结果.总体来看,增加上下文信息后对原帖所反映出的心理健康的判断是有帮助的.对于 SVM 分类模型,在 F³TMH_G 策略选出最优特征组合的基础上增加上下文信息后,*Flagged*、*Urgent*、*All* 的准确率为 0.916、0.727、0.730,较不使用上下文时分别提高了 4.8%、12.2%、3.8%,*Non-green* 上的 *F*1 值也从 0.467 提升到 0.479.同时在所有特征组合 ALLF=(WEF+BAF+NGF+LGF+TPF)的基础上增加上下文信息并使用降噪自编码融合后,其 *Urgent* 的 *F*1 值和 *Acc* 值分别为 0.627 和 0.762 较不使用上下文提高 11.8%和 15.9%,其它指标也普遍有所提升.

表 7 上下文信息对心理健康评估的影响(SVM 分类模型)

特征组合	*Non-green*	*Flagged*		*Urgent*		*All*	
	*F*1	*F*1	*Acc*	*F*1	*Acc*	*F*1	*Acc*
WEF+BAF+NGF+LGF	**0.467**	0.852	0.874	0.571	0.648	0.570	0.703
WEF+BAF+NGF+LGF+CTXF	**0.479**	**0.872**	**0.916**	0.552	0.727	**0.585**	**0.730**
ALLF (F^3TMH_DA)	0.464	0.856	0.858	0.561	0.657	0.567	0.695
ALLF+CTXF (F^3TMH_DA)	0.466	0.858	0.853	**0.627**	**0.762**	0.569	0.705

对数据的进一步分析发现,上下文信息有助于区分那些通过对其他帖子的回复或评论来表达自己态度的帖子.例如样本"@TOM-RO I can't really talk about it coz of the guideline's. It doesn't matter anyway",内容较短,从该贴自身来分析难以准确区分该贴作者此时的心理健康状况,因此仅使用该贴自身信息时模型将其识别为 *amber* 类.在理解其上文信息后发现,该帖是对 TOM-RO 提出的 guideline's 的否定,其情感主要体现在上文中的 guideline's 中,因此在结合该贴的上文信息后模型能较为准确地识别该帖为 *crisis* 类.

4.6 与深度学习方法的比较

相比于本文基于特征工程的方法,也有研究者利用目前较为流行的深度学习方法自动评估在线论坛用户心理健康,取得了很好的效果.

同样在 CLPsych2017 共享任务提供的数据集上,Yates 等人[32]使用卷积神经网络自动学习特征,并采用 Categorical Cross Ent.、MSE、Class Metric、Class Metric(Ordinal)四种损失函数,构建了基于深度学习的在线论坛用户心理健康评估模型,实验结果如表 8 所示.

表 8 基于深度学习的用户心理健康评估模型的评测结果[32]

损失函数	*Non-green*	*Flagged*		*Urgent*		*All*	
	*F*1	*F*1	*Acc*	*F*1	*Acc*	*F*1	*Acc*
Categorical Cross Ent.	**0.37**	0.88	0.90	0.48	0.83	0.50	0.71
MSE	0.31	0.84	0.84	0.54	0.84	0.44	0.64
Class Metric	0.30	0.88	0.89	0.46	0.81	0.45	0.68
Class Metric(Ordinal)	0.34	**0.89**	0.90	0.49	0.81	0.48	0.69

对比表 8 和表 6 以及表 7 中的各项指标,可以看出,采用深度学习的方法在 *Non-green F*1、*Urgent F*1 和 *All F*1 上均与本文的方法有较大的差距.值得注意的是,基于深度学习的方法在 *Flagged F*1 值最高达到 0.89,高于表 7 中的最高值 0.872,说明基于深度学习的方法能较好地区分 *green* 类和非 *green* 类的帖子,但无法较好地划分帖子所反映的心理危机等级.产生这种结果的原因可能与样本数量有一定关系:基于深度学习的方法通过优化网络结构参数来提取特征,虽然可以避免传统方法繁琐的特征工程,但是在小规模数据集上的效果往往不如传统的机器学习方法,而在在线论坛用户心理健康评估这一问题上,心理危机程度高的用户就非常少.

4.7 CLPsych2017 Shared Task 评测结果

表 9 为参加 CLPsych2017 共享评测任务的 16 支队伍最好成绩排名①,其中来自 Xia 和 Liu 的参赛系统采用了基于投票特征融合策略的 F^3TMH_V.从评测结果来看,本文提出的基于多特征融合方法的在线论坛用户心理健康自动评估框架 F^3TMH 有极强的竞争力,其中 F^3TMH_V 在来自 19 家单位的 16 支参赛队伍 251 个参赛系统中取得了第一的成绩,而 F^3TMH_G 的 *Non-green F*1 值也达到了 0.467.由于时间仓促,本文提出的其他特征融合策略没能参赛.

CLPsych2017 没有要求参赛者提供论文来详细介绍所使用的方法,只要求参赛者在提交参赛结果时对使用的方法作简要说明,因此,其他参赛队伍所使用的方法只能从 CLPsych2017 workshop 的会议报告②中简要了解到.例如,Altszyler 除了利用帖子本身的信息外,还利用了该帖子作者发布的其它帖子,模型使用的特征包括 N-Grams、词向量、发帖时间、用户关注和转发等;Nair 等人使用帖子的文档向量(Doc2Vec)、主题、句法等特征构建 LSTM 深度学习模型;Yates 等人[32]基于卷积神经网络模型,除使用帖子自身内容外,还利用了帖子所在主题

① https://drive.google.com/file/d/0BwDHn92ZGZW9Q1liWUw5UDZxVFE/view,该排名是以每支参赛队伍提交的系统结果的 *Non-green F*1 值为依据

② https://drive.google.com/file/d/0BwDHn92ZGZW9Rzl0OFpsajY5MGc/view

表 9　CLPsych2017 shared task 评测结果

Participants	*Non-green*	*Flagged*		*Urgent*		*All*	
	*F*1	*F*1	*Acc*	*F*1	*Acc*	*F*1	*Acc*
Xia and Liu	**0.467**	0.845	0.877	0.591	0.746	—	**0.708**
Altszyler	0.462	0.881	0.883	0.598	0.825	—	0.695
Nair et al.	0.461	0.800	0.820	0.571	0.843	—	0.678
Han et al.	0.447	0.843	0.858	0.571	0.798	—	0.683
Qadir et al.	0.436	0.757	0.803	0.529	0.840	—	0.683
Gamaarachchige et al.	0.413	0.862	0.868	0.585	0.830	—	0.678
French et al.	0.391	0.877	0.883	0.550	0.788	—	0.668
Vajjala	0.388	0.765	0.798	0.599	0.823	—	0.643
Miftahutdinov et al.	0.373	0.870	0.825	0.497	0.808	—	0.653
Hoenen	0.324	0.724	0.710	0.407	0.753	—	0.523
Yates et al.	0.319	0.867	0.880	0.488	0.843	—	0.700
Patra and Kirk	0.290	0.719	0.778	0.420	0.800	—	0.638
Kennington & Mehrpouyan	0.281	0.806	0.825	0.330	0.808	—	0.650
Desmet and Jacobs	0.219	0.606	0.513	0.363	0.535	—	0.275
Rose and Bex	0.187	0.627	0.623	0.213	0.723	—	0.443
Morales and Levitan	0.086	0.426	0.630	0.004	0.780	—	0.540

块(Thread)中其它帖子的内容；Han 等人尝试了深度神经网络、线性 SVM 等方法.

总的来看，参赛者使用的评估模型以及用于评估模型的特征丰富多彩、各有亮点，对推动心理健康自动评估的研究有积极意义. 然而，从评测结果来看，最好参赛系统的 *Non-green* *F*1 值仍不到 0.5，说明在线论坛用户心理健康自动评估问题是一个较难的问题，还有很大的提升空间.

5　总结与展望

本文针对在线论坛等社会网络上的用户帖子，提出了基于多特征融合的在线心理健康自动评估框架 F^{3}TMH，自动评估帖子所反映出的用户心理健康危机程度. F^{3}TMH 利用了用户的行为表现和帖子自身的属性、帖子的语言或用词风格、帖子内容(包括 N-Grams、主题、词向量)、上下文等多种类型的特征，并通过贪婪法、投票法、后期融合法以及基于深度学习的降噪自编码法四种不同的策略对特征进行融合.

基于 CLPsych2017 shared task 任务，考察了各类特征以及不同的特征融合策略对心理健康自动评估性能的影响. 实验结果表明，相对于行为与属性特征(BAF)、语言特征(LGF)、主题特征(TPF)和 N-Grams 特征(NGF)，基于 Word2Vec 的词向量特征 WEF 在透过内容较少的论坛帖子自动评估在线论坛用户心理健康危机程度的问题上表现更好. 实验还发现，投票法特征融合策略 F^{3}TMH_V 在识别心理健康危机程度较高的 *crisis* 类和 *red* 类帖子时更有效，而且通过分类器代替投票器的后期特征融合策略 F^{3}TMH_L 能进一步提升 *crisis* 类帖子的识别效果及总体效果. 降噪自编码特征融合策略 F^{3}TMH_DA 能在有效降低维度的同时，能更好地提升样本数量较多类型帖子(*green* 和 *amber*)的识别效果. 在考察原帖上下文信息后发现，上下文信息有助于理解原帖内容，对识别原帖类别有辅助作用. 尽管本文提出的四种特征融合策略没有全部参加 CLPsych2017 shared task 评测，但是 F^{3}TMH_L 和 F^{3}TMH_DA 都展现出了明显的优势.

通过实验分析，我们发现可以从以下几个方面对模型进行改进：

(1)充分挖掘帖子作者的行为模式和用户属性. 本文只考虑了单个帖子的相关信息，不能精准刻画作者的行为模式和用户属性. 有必要从作者发布的全部帖子的行为，乃至行为的变化上，挖掘用户行为模式或用户属性.

(2)充分利用未标注的数据. 人工标注一方面要求标注者有一定的专业知识，另一方面标注的工作量大，标注的样本数量有限，需要采用半监督学习方法或是标签传播等方法扩充训练样本，更有效地利用未标注的数据.

(3)结合深度学习方法与特征工程. 基于传统特征工程的机器学习方法虽然表现不错，但是依然有一些特征可能会被忽视. 结合深度学习方法，并在其基础上引入传统特征工程的方法，如：采用注意力机制学习训练，这是我们将来的工作之一.

致　谢　由衷感谢审稿专家给出的修改意见！

参　考　文　献

[1]　Liao Cheng-Ju, Feng Zheng-Zhi. Mechanism of affective and

cognitive-control brain regions in depression. Advances in Psychological Science, 2010, 18(2): 282-287(in Chinese)
(廖成菊, 冯正直. 抑郁症情绪加工与认知控制的脑机制. 心理科学进展, 2010, 18(2): 282-287)

[2] Liang X, Gu Siqi, Deng J, et al. Investigation of college students' mental health status via semantic analysis of Sina microblog. Wuhan University Journal of Natural Sciences, 2015, 20(2): 159-164

[3] McGuire S. U.S. Department of agriculture and U.S. department of health and human services, dietary guidelines for Americans. Advances in Nutrition, 2011, 2(3): 293-294

[4] Ferrari A J, Norman R E, Freedman G, et al. The burden attributable to mental and substance use disorders as risk factors for suicide: Findings from the global burden of disease study 2010. PLoS One, 2014, 9(4): e91936

[5] Almeida H, Queudot M, Meurs M J. Automatic triage of mental health online forum posts CLPsych 2016 system//Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality. San Diego, USA, 2016: 183-187

[6] Rey-Villamizar N, Shrestha P, Solorio T, Pedersen T. A semi-supervised approach for the CLPsych 2016 shared task//Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality. San Diego, USA, 2016: 171-176

[7] Brew C. Classifying ReachOut posts with a radial basis function SVM//Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality. San Diego, USA, 2016: 138-142

[8] Cohan A, Young S, Goharian N. Triaging mental health forum posts//Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality. San Diego, USA, 2016: 143-147

[9] Kim M S, Wang Y, Wan S, et. al. Data61-CSIRO systems at the CLPsych 2016 shared task//Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality. San Diego, USA, 2016: 128-132

[10] Desmet B, Jacobs G, Hoste V. Mental distress detection and triage in forum posts?: The LT3 CLPsych 2016 shared task system//Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality. San Diego, USA, 2016: 148-152

[11] Malmasi S, Zampieri M, Dras M. Predicting post severity in mental health forums//Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality. San Diego, USA, 2016: 133-137

[12] Friedenberg M, Amiri H, Iii H D, et al. The UMD CLPsych 2016 shared task system: Text representation for predicting triage of forum posts about mental health//Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality. San Diego, USA, 2016: 158-161

[13] Tsugawa S, Mogi Y, Kikuchi Y, et al. On estimating depressive tendencies of Twitter users utilizing their tweet data//Proceedings of the IEEE Virtual Real. Lake Buena Vista, USA, 2013: 1-4

[14] Choudhury M D, De S. Mental health discourse on reddit: Self-disclosure, social support, and anonymity//Proceedings of the 8th International AAAI Conference on Weblogs and Social Media. Ann Arbor, USA, 2014: 71-80

[15] Guan Li, Hao Bi-Bo, Cheng Qi-Jin, et. al. Behavioral and linguistic characteristics of microblog users with various suicide ideation level: An explanatory study. Chinese Journal of Public Health, 2015, 31(3): 349-352(in Chinese)
(管理, 郝碧波, 程绮瑾等. 不同自杀可能性微博用户行为和语言特征差异解释性研究. 中国公共卫生, 2015, 31(3): 349-352)

[16] Guan Li, Hao Bi-Bo, Liu Tian-Li, et. al. A pilot study of differences in behavioral and linguistic characteristics between Sina suicide microblog users and Sina microblog users without suicide idea. Chinese Journal of Epidemiology, 2015, 36(5): 421-425(in Chinese)
(管理, 郝碧波, 刘天俐等. 新浪微博用户中自杀死亡和无自杀意念者特征差异的研究. 中华流行病学杂志, 2015, 36(5): 421-425)

[17] Schrammel J, Tscheligi M. Personality traits, usage patterns and information disclosure in online communities//Proceedings of the British HCI Group Conference on People and Computers: Celebrating People and Technology, British Computer Society. Cambridge, UK, 2009: 169-174

[18] Ross C, Orr E S, Sisic M, et al. Personality and motivations associated with Facebook use. Computers in Human Behavior, 2009, 25(2): 578-586

[19] Bai Shuo-Tian, Hao Bi-Bo, Li Ang, et al. Depression and anxiety prediction on microblogs. Molecular Microbiology, 2014, 5(8): 814-820

[20] Conner K R, Bohnert A S, Mccarthy J F, et al. Mental disorder comorbidity and suicide among 2.96 million men receiving care in the Veterans Health Administration health system. Journal of Abnormal Psychology, 2013, 122(1): 256-263

[21] Choudhury M D, Gamon M, Counts S, et al. Predicting depression via social media//Proceedings of the 7th International AAAI Conference on Weblogs and Social Media. Boston, USA, 2013: 128-137

[22] Choudhury M D, Counts S, Horvitz E. Social media as a measurement tool of depression in populations//Proceedings of the 5th Annual ACM Web Science Conference. Paris, France, 2013: 47-56

[23] Gosling S D, Gaddis S, Vazire S. Personality impressions based on Facebook profiles//Proceedings of the International Conference on Weblogs and Social Media. Boulder, USA, 2007: 26-28

[24] Evans D C, Gosling S D, Carroll A. What elements of an online social networking profile predict target-rater agreement

in personality impressions//Proceedings of the 2nd International Conference on Weblogs and Social Media. Seattle, USA, 2008: 45-50

[25] Wei H, Zhang F, Yuan N J, et al. Beyond the words: Predicting user personality from heterogeneous information//Proceedings of the 10th ACM International Conference on Web Search and Data Mining. Cambridge, UK, 2017: 305-314

[26] Wang X, Zhang C, Ji Y, et al. A depression detection model based on sentiment analysis in microblog social network//Proceedings of the 17th Pacific-Asia Conference on Knowledge Discovery and Data Mining. Gold Coast, Australia, 2013: 201-213

[27] Liu X, Zhu T. Deep learning for constructing microblog behavior representation to identify social media user's personality. PeerJ Computer Science, 2016, 2016(2): e81

[28] Huang X, Li X, Zhang L, et al. Topic model for identifying suicidal ideation in Chinese microblog//Proceedings of the 29th Pacific-Asia Conference on Language, Information and Computation. Shanghai, China, 2015: 553-562

[29] Zhang L, Huang X, Liu T, et al. Using linguistic features to estimate suicide probability of Chinese microblog users//Proceedings of the International Conference on Human Centered Computing. Phnom Penh, Cambodia, 2014: 549-559

[30] Wang X, Zhang C, Sun L. An improved model for depression detection in microblog social network//Proceedings of the IEEE 13th International Conference on Data Mining Workshop. Texas, USA, 2013: 80-87

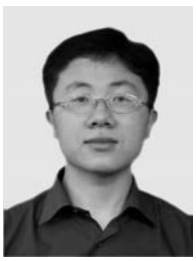
[31] Choudhury D M, Counts S, Horvitz E J, et al. Characterizing and predicting postpartum depression from shared Facebook data//Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing. Maryland, USA, 2014: 626-638

[32] Yates A, Cohan A, Goharian N. Depression and self-harm risk assessment in online forums//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark, 2017: 2968-2978

[33] Manikonda L, Choudhury M D. Modeling and understanding visual attributes of mental health disclosures in social media//Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. Colorado, USA, 2017: 170-181

[34] Lv M, Li A, Liu T, et al. Creating a Chinese suicide dictionary for identifying suicide risk on social media. PeerJ, 2015, 3(10.7177): e1455-e1469

[35] Stirman S W, Pennebaker J W. Word use in the poetry of suicidal and nonsuicidal poets. Psychosomatic Medicine, 2001, 63(4): 517-522



LIU De-Xi, Ph.D., professor, Ph.D. supervisor. His research interests include social media processing, information retrieval, and natural language processing.

XIA Xian-Yi, M.S. candidate. His research interest is social media processing.

WAN Chang-Xuan, Ph.D., professor. His research interests include Web data management and sentiment analysis.

LIU Xi-Ping, Ph.D., associate professor. His current research interest is Web data management.

JIANG Teng-Jiao, Ph.D., lecturer. Her current research interest is sentiment analysis.

FU Qi, Ph.D. candidate. Her research interest is social media processing.

Background

Automatic assessment of mental health of social network users has wide and important application in the field of natural language processing and psychology, which has attracted wide attentions from researchers in computer science and psychology. Social network is often used by people to express their opinions, mental health problems and even turn to/for help on social network. Social network has being become a significant resource to study textual language related to depression, self-harm, suicide, etc. Most related works in this area used various classifiers based on feature engineering such as bag-of-words, topic model, doc2vec, a stack of feature-rich random forest, support vector machine and deep learning (e.g. convolution neural network, long short-term memory) etc. Since 2014, the Computational Linguistics and Clinical Psychology Workshops released a shared task to explore the approaches for detecting the mental health of uses on forums.

This team has done some works about sentiment analysis and psychology analysis on financial text and social network text, which were published on or accepted by Chinese Journal of Computers and Journal of Chinese Information Processing.

This work is supported by the National Natural Science Foundation of China (Nos. 61762042, 61363039, 61562032), the Transformation Project of Scientific and Technological Achievements from Universities in Jiangxi Province (No. KJLD14035), and the Natural Science Foundation of Jiangxi Province (Nos. 20171BAB202021, 20152ACB20003). These projects focus on mental health analysis for social network users, multi-microblogs summarization, sentiment summarization.