Beijing Forest Studio

北京理工大学信息系统及安全对抗实验中心

认知扭曲识别研究

博士研究生 陈星星

2026年04月06日

问题回溯

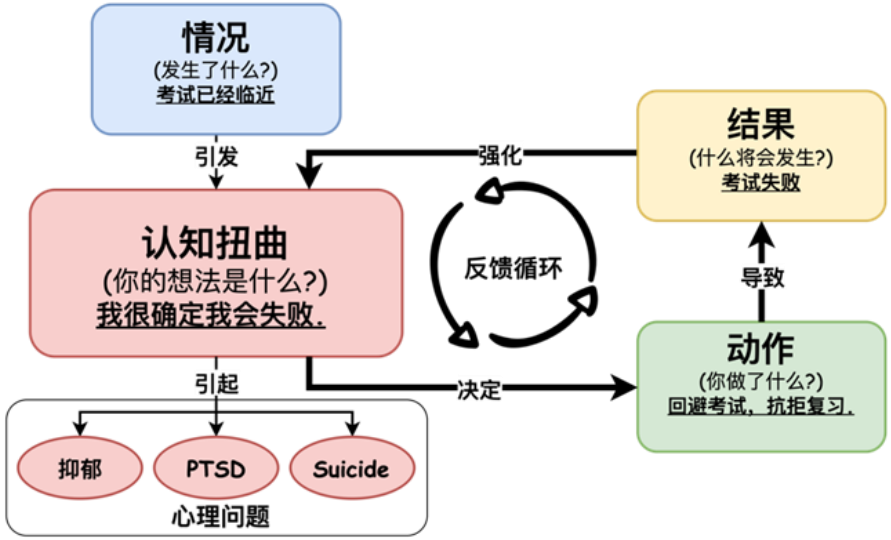
- 总结反思
  - 时间把控不足，第二个算法讲解过快
  - 内容讲解不够深入

内容提要

- 预期收获
- 题目内涵解析
- 研究背景与意义
- 研究历史与现状
- 知识基础
- 算法原理
  - Multi-View Attention Multiple-Instance Learning Enhanced by LLM Reasoning for Cognitive Distortion Detection
  - Enhancing Depression Detection with Cognitive Distortion and User-level Information
- 特点总结与工作展望
- 参考文献

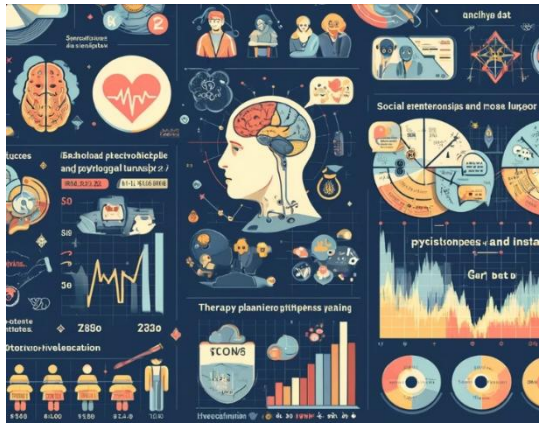
预期收获

- 预期收获
  - 了解认知扭曲的基本概念与主要类型
  - 理解认知扭曲识别基本原理
  - 掌握认知扭曲驱动下游任务的方式
- 研究背景
  - 抑郁、焦虑等心理健康问题已成为重要公共卫生问题，早期识别与及时干预具有重要现实意义
  - 认知扭曲是认知行为治疗中的核心概念，与抑郁、焦虑等心理障碍的形成和维持密切相关
  - 现有很多研究主要关注情绪倾向或表层语义，但难以刻画个体更深层的思维偏差与认知模式
  - 认知扭曲往往存在表达隐晦、类别重叠、多种扭曲共现和上下文依赖强等特点，给自动识别带来困难



目标内涵—认知扭曲识别研究

- 研究目标
  - 更准确、更细粒度地识别文本中的认知扭曲类型
  - 将认知扭曲作为关键心理特征，引入用户级抑郁检测任务
  - 增强模型的解释性与应用价值



- 内涵解析
  - 心理健康计算：运用自然语言处理、机器学习和大语言模型等方法，对抑郁、焦虑、认知偏差等心理健康相关现象进行自动分析、识别与建模
  - 认知扭曲识别：从文本中自动识别个体是否存在以及属于哪一类认知扭曲，如非黑即白、过度概括、情绪化推理和贴标签等

研究历史

2017

Simms等用机器学习文本分析来检测认知扭曲，证明了认知扭曲可以被视为一个自然语言处理任务来处理，但整体上仍处于方法验证和任务定义的早期阶段

2020

Shickel等提出了一个较为系统的认知扭曲检测框架，将任务划分为扭曲/非扭曲二分类和具体扭曲类型多分类两个层次，并在众包文本和真实在线治疗文本上对15类常见认知扭曲进行检测与分类

2021

Shreevastava和Foltz在患者治疗师交互文本场景，发现基于预训练句向量和SVM的方法效果较好；把认知扭曲识别从一般文本分类推进到更接近真实治疗语境的心理健康文本分析

2022

Lybarger等指出认知扭曲往往具有明显的对话上下文依赖性，因此在患者—治疗师短信交流中引入动态多轮上下文建模，并采用BERT-based架构表示会话消息，探索对话历史对扭曲预测的作用

2022

Ding等进一步聚焦认知扭曲识别中的低频类别与类别不平衡问题，比较了数据增强方法与领域专用预训练模型两条路线，发现它们都能在一定程度上改善稀有扭曲类别的识别效果

2023

Chen等将大语言模型引入认知扭曲检测，提出思想诊断提示(DoT)，通过主观性判断、对比推理和schema分析等分步提示，引导模型进行更接近心理治疗推理过程的判断

2024

Elsharawi和El Bolock提出C-Journal，在众包数据集基础上比较Naïve Bayes、CNN、LSTM和DNN等方法，并将认知扭曲分类能力嵌入日记应用场景

2025

Kim等提出一种结合LLM推理与多实例学习(MIL)的认知扭曲检测框架，将一句话拆分为三个心理维度，并由多个LLM生成多个扭曲实例，再通过多视角门控注意力进行聚合分类

2026

Wan等将认知扭曲进一步作为用户级抑郁检测的关键中间特征，提出融合多粒度认知扭曲学习、帖子级扭曲感知机制和用户级行为信息的抑郁检测模型，

知识基础 认知扭曲

- 认知扭曲
  - 心理学家阿伦.贝克认为：认知扭曲（cognitive distortion）是一种思维的错误，它造成了人类处理信息过程的困难，最终导致了心理障碍
    - 个体在理解和解释现实事件时出现的系统性思维偏差，常导致不符合事实的负性结论
  - 常见类型

认知扭曲类型	核心含义	典型表现
非黑即白思维	把事情看成两个极端，没有中间地带	要么完美，要么彻底失败
过度概括	根据一次事件或有限证据，推导出普遍结论	这次失败了，说明我以后都不行
心理过滤	只关注负面部分，忽略正面信息	别人夸了我，但我只记住那个小错误
否定积极面	否认正向经历或表扬的价值，认为“不算数”	他们说我做得好，只是客气而已
预言式推断	预设未来会朝坏方向发展	我肯定会失败，事情一定会变糟
读心术	主观揣测别人怎么想，通常是负面推断	他们肯定觉得我很差劲
个人化归责	将他人的负面行为归咎于自己，没有合理的联系	我的老板生气了，一定是我做错了什么

知识基础 认知扭曲识别

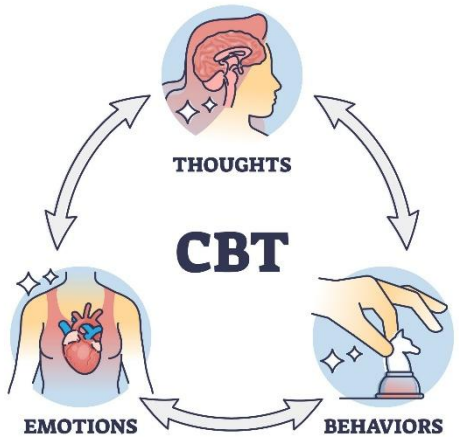
- 认知扭曲检测
  - 输入一句话或一段心理相关文本，识别其对应的认知扭曲类型
- 难点
  - 一句话可能含有多个扭曲
  - 扭曲类型之间语义相近
  - 表达具有强主观性和歧义性

Original: 我好累，真的好累。考试永远过不了，我不能拿我前途去赌，我好害怕。 Translation: Feeling so drained. Exams never seem to end and I can't gamble with my future. Seriously scared.
Ground truth labels: Over-generalization
GPT-4 predictions: The fortune teller error, Should statements
Original: 饭，今天是 5 月了。我好像很烦很迷茫。我现在又觉得世界不过如此，我不想走这一遭了。我马上高考了，想到父母心真的真的会很痛。拜托、拜托，真的太难坚持了，我想睡了。 Translation: Fan, it's already May. Feeling so overwhelmed and lost. The world seems so meh right now. I can't bear the thought of going through this, especially with college entrance exams coming up. Thinking of my parents just hurts so much. Please, it's really hard to keep going. I just wanna sleep.
Ground truth labels: Mental filter
GPT-4 predictions: Mental filter, Disqualifying the positive, The fortune teller error, Blaming oneself

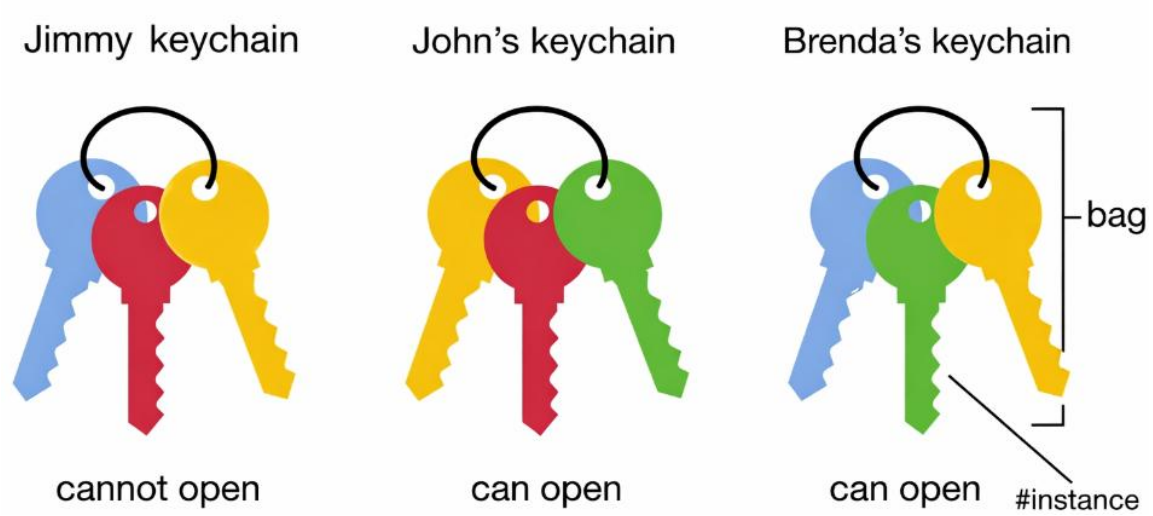
Multi-View Attention Multiple-Instance Learning Enhanced by LLM Reasoning for Cognitive Distortion Detection

知识基础　认知行为疗法

- 认知行为疗法（CBT）
  - CBT的理论基础建立在认知理论之上，特别是贝克发展的认知模型
    - 个体的情绪反应并非直接由事件决定，而是由其对事件的解释、信念和自动想法所中介
    - 通过对个体认知层面（思维模式），以及行为层面（行动方式的改变），达成缓解情绪困扰、优化心理功能的目的
- 认知三角
  - 人的情绪（emotion）—认知/想法（thought）—行为（behavior）是彼此影响的



- 多实例学习
  - 一种弱监督学习框架，将数据组织成一种层级结构：由多个实例构成一个包（Bag），标签是在“包”的层面上给出的
  - 在多实例学习中，每个实例可能都有各自的特点，但包只会有一个总体标签，告诉你这个包里是否有“重要”的东西
    - 阳性包：包里至少有一个实例是“阳性”的，比如它含有某种关键特征
    - 阴性包：包里的所有实例都是“阴性”的，也就是说它们都没有关键特征



ELB TIPO

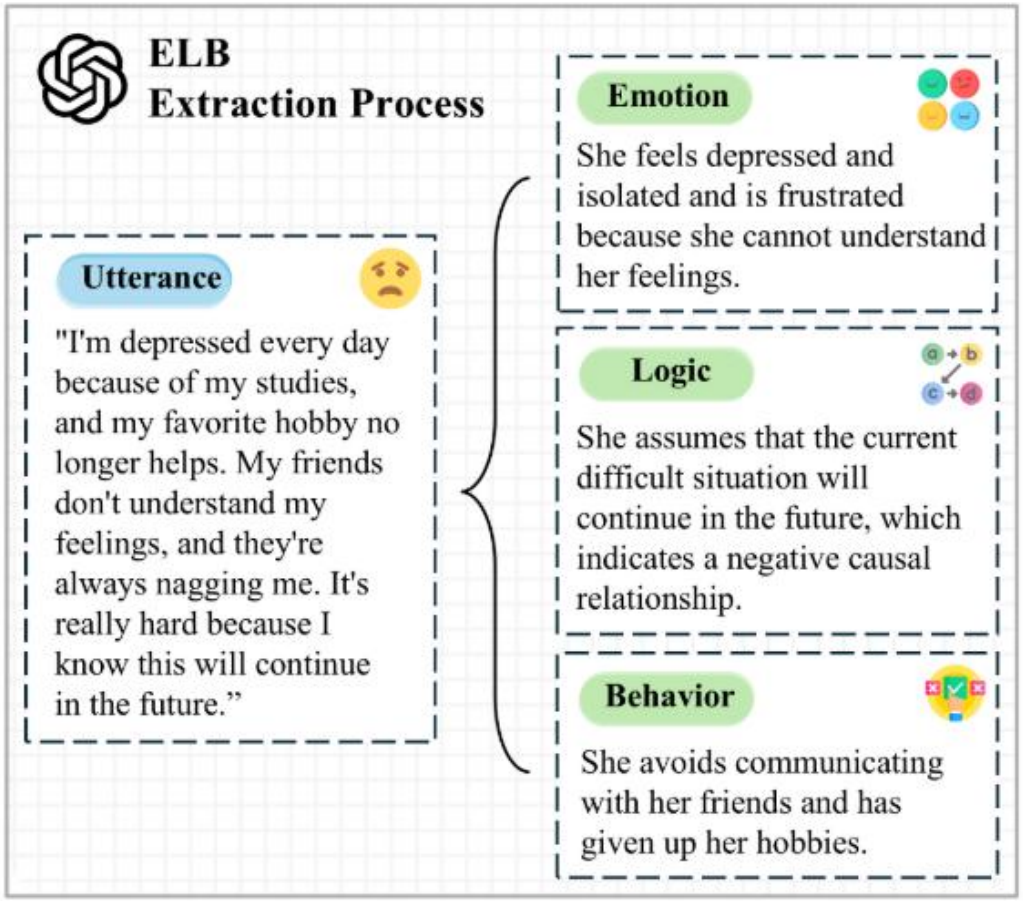
T	目标	构建更细粒度、可解释、能处理多认知扭曲共现的自动认知扭曲检测框架
I	输入	KoACD数据集：4510条单标签认知扭曲样本。Therapist QA数据集：1597条多标签认知扭曲的样本
P	处理	1.结合LLM与多实例学习（MIL）架构，增强可解释性和表达推理能力 2.话语被分解为情绪、逻辑和行为组件，LLMs处理推断多个扭曲实例，每个实例具有预测类型、表达式和模型分配的显著性分数 3.实例通过多视角门控注意力机制进行最终分类
O	输出	认知扭曲类型

P	问题	1.一个句子里可能同时有多个认知扭曲 2. 认知扭曲识别缺乏可解释性
C	条件	LLM能较可靠地抽取文本中情绪、逻辑、行为并生成认知扭曲实例
D	难点	1.不同认知扭曲之间语义相似、边界模糊 2.扭曲依赖隐含的情绪、逻辑和行为线索，导致容易漏检和混淆
L	水平	2025 预印本

ELB 情感-逻辑-行为抽取

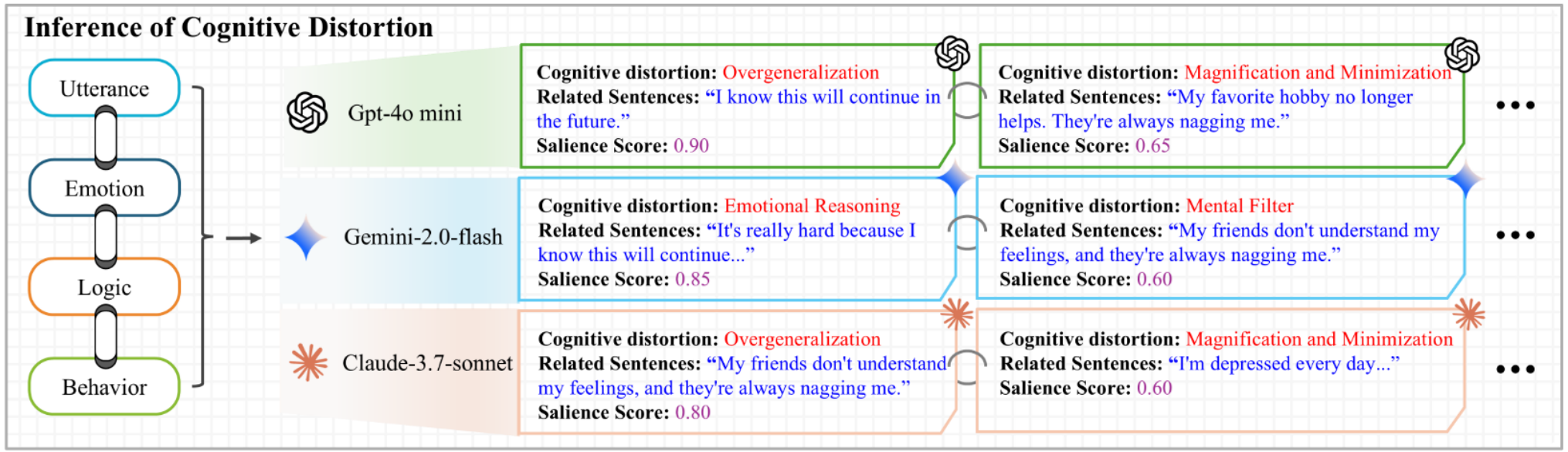
- 把原始句子拆成ELB三个心理维度
  - Emotion：情绪
  - Logic：思维逻辑
  - Behavior：行为意图
- 情绪表达：我很难过；我很焦虑
- 逻辑推断：既然失败一次，就说明我永远不行
- 行为倾向：我决定再也不去尝试了

给下游认知扭曲推理提供更显式的心理线索
提高对模糊、重叠扭曲的识别能力
提升可解释性



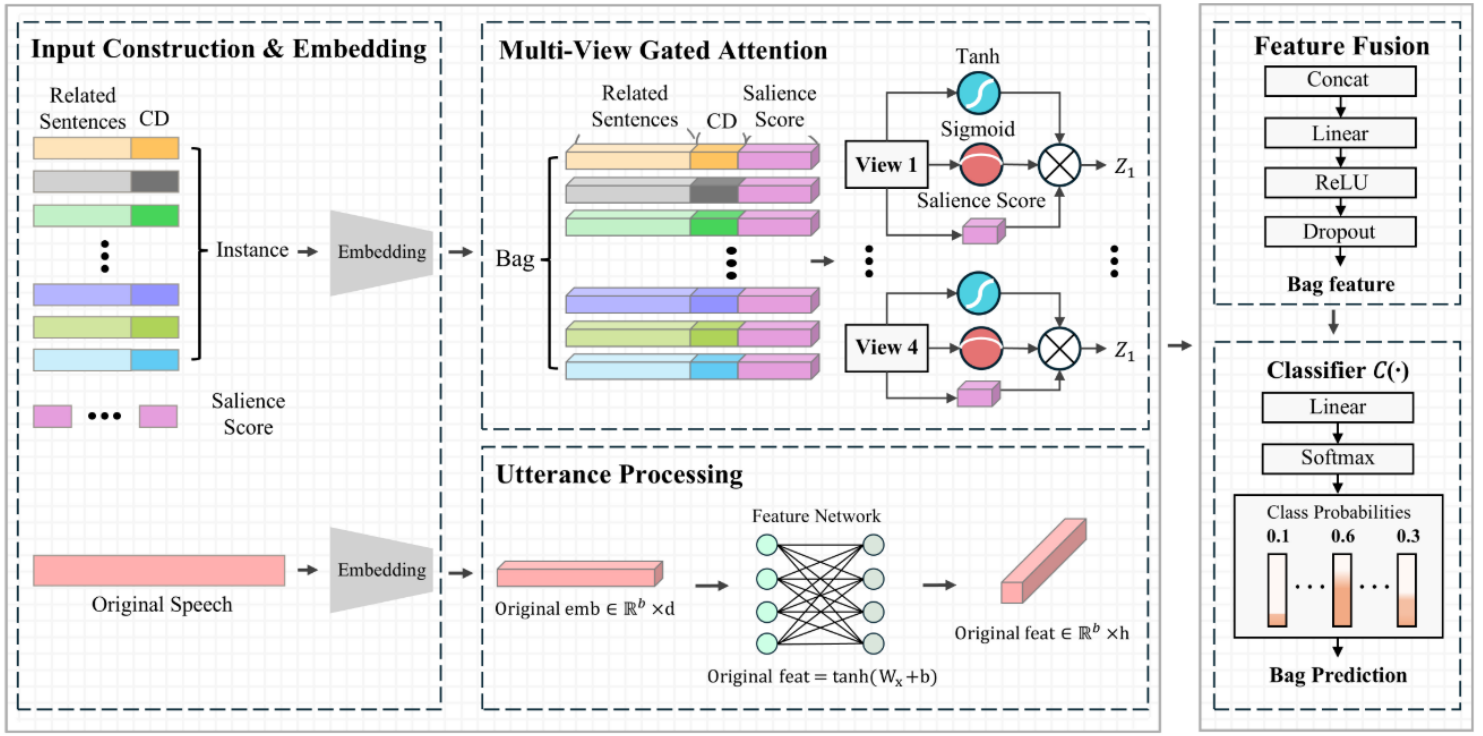
ELB 基于大语言模型的认知扭曲实例推断

- 多个 LLM 为同一句话提出多个认知扭曲实例
  - GPT-4o、Gemini 2.0 Flash、Claude 3.7 Sonnet
- 基于同一个句子和 ELB 结构化信息，各自输出若干条“认知扭曲实例”。每个实例包含三项：
  - 预测的认知扭曲类型、对应的相关文本片段、该实例的显著性得分



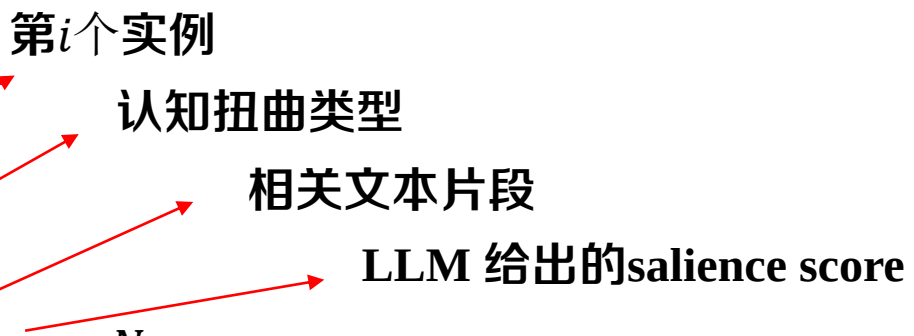
ELB MIL建模和包构建

- MIL模型对实例进行包级分类
  - 一个原始表达看成一个 包（bag）
  - 原始表达里由LLM推理出的多个认知扭曲表达，看成多个实例
- 利用多视角门控注意力的MIL模型，把多个实例加权聚合，最后输出认知扭曲类别



ELB MIL建模和包构建

- 把每一个实例集合当作一个包（bag）



$$B = \{x_i = (type_i, text_i, s_i)\}_{i=1}^{N}$$

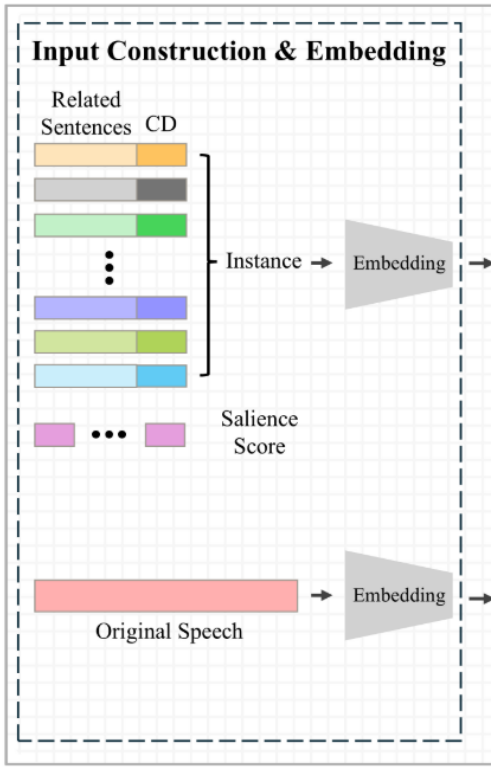
- 一个bag不是一整句文本，而是一组由LLM发现的扭曲表达单元
- 实例级显著性评分作为基于注意力整合结构中的加权信号
  - LLM给出的显著性得分先做归一化：

$$\hat{p}_i = \frac{s_i}{\sum_{j=1}^{N} s_j}$$

- 把每个实例的显著性转成当前bag内部的相对权重

- 原句与实例嵌入
  - 原句嵌入：原始表达被编码为一个384维向量
    - 当实例没有覆盖正确标签时，原句信息仍然可提供补救
    - 保留一些实例级表示没有覆盖到的全局心理和上下文信息
  - 实例嵌入：实例由“预测类型 + 相关文本”拼接后编码为384维向量
    - 每个实例同时包含语义内容和模型赋予的类别信息

MIL模型的最终输入是结合句子嵌入和所有实例嵌入的序列



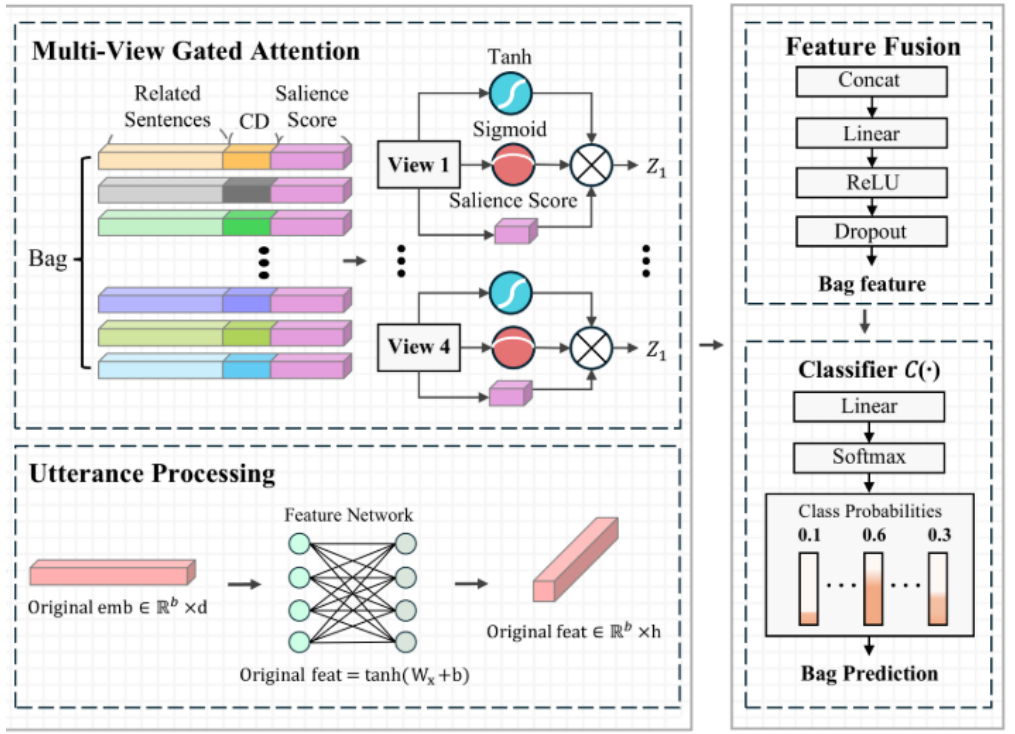
ELB 多视角门控注意力

- 实例注意力得分

$$h_i = \sigma(W_g \cdot x_i) \cdot \tanh(W_f \cdot x_i) \cdot s_i$$

  - $\tanh(W_f \cdot x_i)$提取实例的语义特征
  - $\sigma(W_g \cdot x_i)$决定哪些特征通道要被放大或抑制
  - 把实例重要性判断融入注意力
- 多视角实例建模
  - 让注意力从不同“观察角度”去看包内的实例

$$h_{multi} = \frac{1}{K}\sum_{K=1}^{K} h^{(k)}$$



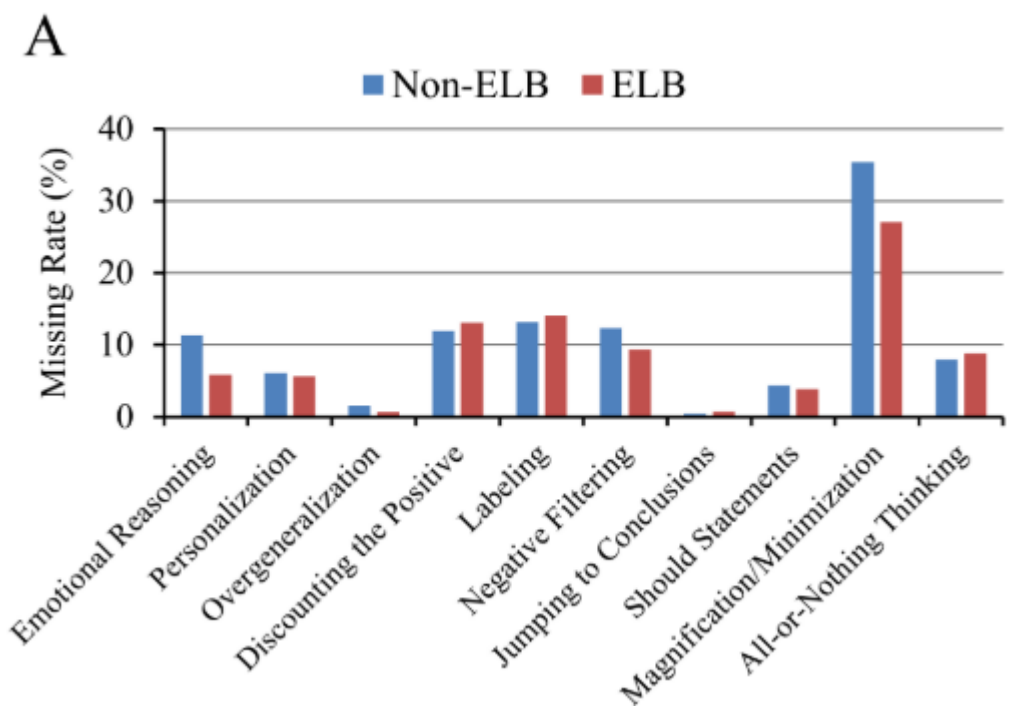
- 原句特征变换与融合
  - 原始句子向量z先做一个非线性投影；再与多视角实例聚合表示h_{multi}拼接再利用Softmax分类器，输出认知扭曲类别

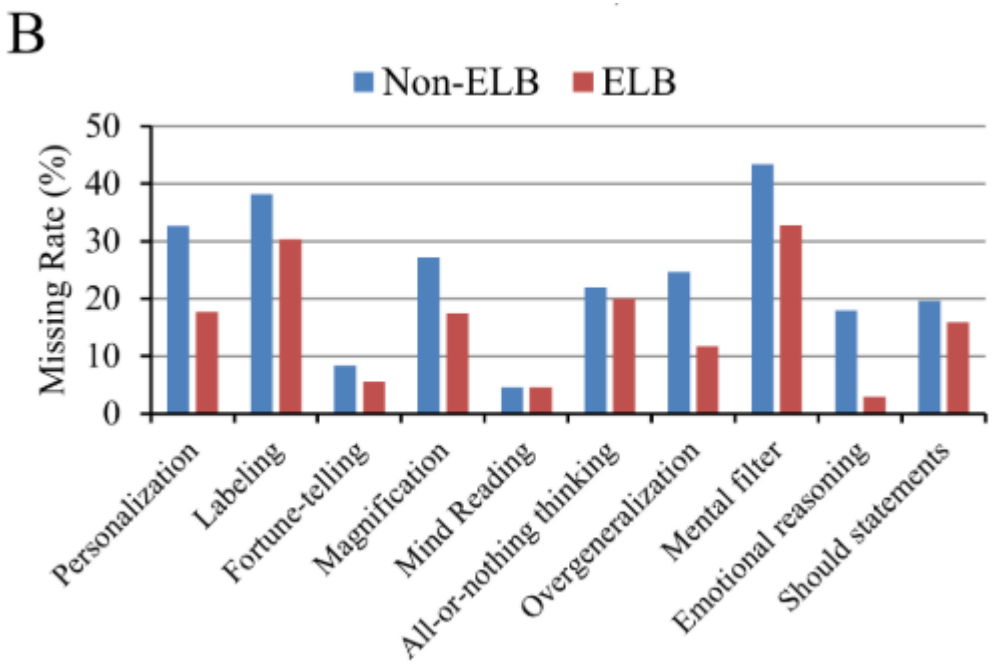
实验设计　数据资源

- 数据集
  - KoACD数据集：来源于NAVER Knowledge iN平台上的咨询文本。抽取5000条原始表达，并且每类先采样500条，最终保留 4510条单标签样本。
  - Therapist QA数据集：来源于Kaggle上公开的patient-therapist异步问答日志。使用了一个精炼后的英文子集，共1597条原始表达，只保留有认知扭曲的样本，最多可对应两个扭曲类型
- 评价指标
  - Missing Rate：如果LLM生成的所有实例里，没有一个实例包含当前原始表达的真值，就记为missing。衡量上游实例生成是否覆盖正确标签
  - Weighted F1：把每个类别的样本数量作为权重，计算加权F1

实验设计　对比实验

- **Missing Rate**
  - **KoACD：加入 ELB 后，大多数认知扭曲类型的missing rate都下降了**
  - **Therapist QA：在英文数据集上，改善更明显**
  - **ELB结构化输入能显著降低正确认知扭曲标签的漏检率**
- **ELB最核心的价值是：帮助LLM更准确地想到正确的认知扭曲实例**





实验设计　对比实验

- **Weighted F1**
  - **在KoACD和Therapist QA上ELB 与显著性得分取得最佳分类性能，说明先增强候选实例覆盖，再强化关键实例加权是有效的**

Methods	KoACD		Therapist QA	
	Val F1	Test F1	Val F1	Test F1
Baseline	0.504 ± 0.019	0.473 ± 0.015	0.410 ± 0.038	0.340 ± 0.037
ELB	0.519 ± 0.016	0.483 ± 0.017	0.438 ± 0.028	0.378 ± 0.036
Salience	0.518 ± 0.015	0.486 ± 0.014	0.428 ± 0.036	0.360 ± 0.035
ELB + Salience	**0.529 ± 0.018**	**0.505 ± 0.014**	**0.460 ± 0.029**	**0.394 ± 0.034**

Dataset	Utterance
KoACD	*I did well in the club interview, but I feel crushed by the thought that I must live up to everyone's expectations. Like the president said, if I don't get selected, does that mean I'm not qualified to be a leader? Do I have to be perfect at everything? I'm definitely lacking. I must do better.* (동아리 면접을 봤다. 발표는 잘했지만, 모두의 기대에 부응해야 한다는 생각에 짓눌린다. 회장 언니 말처럼 뽑히지 못하면 나는 리더가 될 자격이 없는 사람일까? 모든 걸 완벽하게 해내야만 하는 건가? 내가 부족한 건 분명하다. 나는 더 잘해야만 해.)
Therapist QA	*I have been suffering from bulimia for four months now. I realize the health risks and I know I have a problem. I have been trying to stop for a month now with no success. Before this problem I was healthy and now I fear that all my hard work I have completed over the years to be a healthy person are going down the drain. To be honest I am not sure what started my ED, but my main focus is to overcome it. I know that I have some self esteem issues and I will continue to work on that, but do you have any advice or tricks to stop these behaviors that have seemed to become habitual and uncontrollable. I know that getting professional help is probably the best way to go, but that is not me. I have always dealt with my problems in the past and I would like to give this a shot. So if you have any suggestions or tips to help me slowly stop these bulimic behaviors I would appreciate it so much.*

算法总结 ELB

- 算法贡献
  - 引入ELB结构化表示：把话语拆成情绪-逻辑-行为，增强心理结构建模与可解释性
  - 提出LLM + MIL框架：将一句话中的多个认知扭曲表达建模为多个实例，再做包级分类，更适合处理扭曲共现
  - 融合显著性与多视角注意力：将LLM显著性分数纳入多视角门控注意力，提高关键实例聚合效果

- 算法不足
  - 依赖上游LLM实例生成质量：一旦正确标签未被召回，下游MIL很难弥补
  - 实例级监督与平衡性不足：实例没有真值标注，且不同类型实例分布不均，容易影响注意力学习

Enhancing Depression Detection with Cognitive Distortion and User-level Information

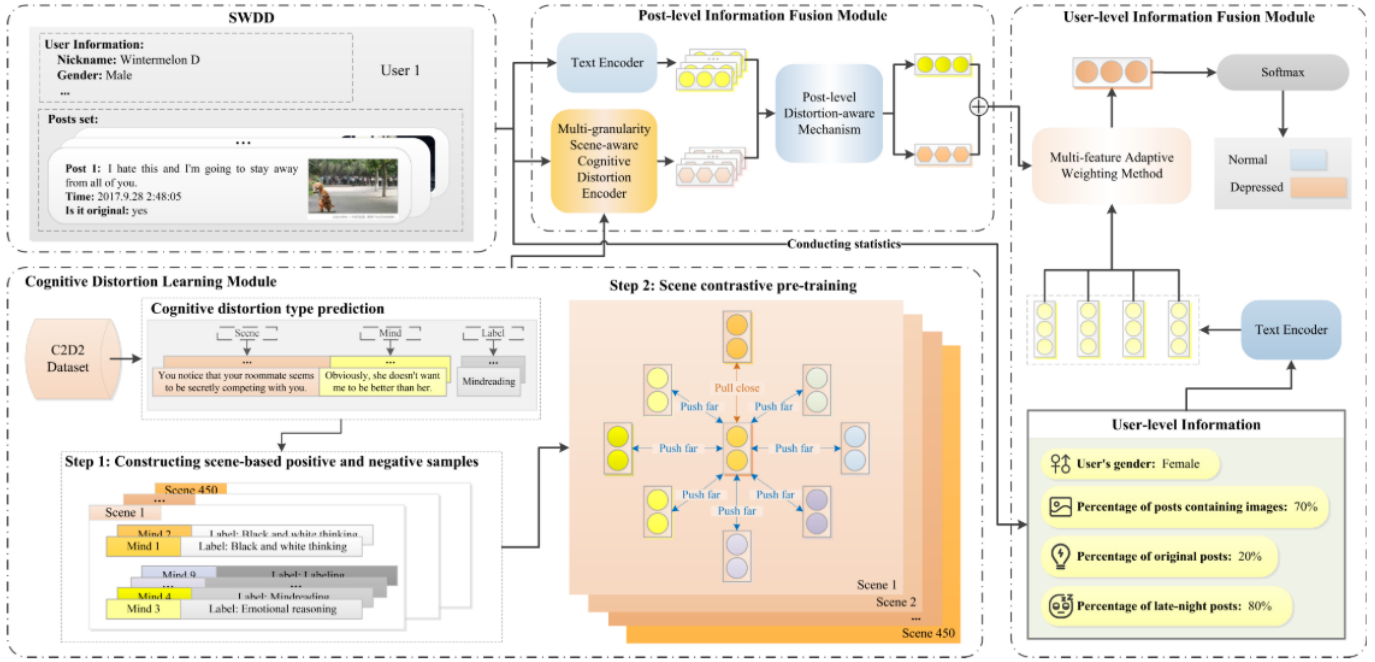
2DM-CDUI TIPO

T	目标	构建融合认知扭曲和用户层信息的抑郁检测模型
I	输入	社交媒体帖子数据集，SWDD-T1：3600个样本、SWDD-T2：7400个样本，一个用户的一组帖子和部分个人信息
P	处理	1.采用多粒度认知扭曲学习方法捕获用户文本中潜在的多粒度认知扭曲信息 2.采用后级失真感知机制识别与用户主观认知相关的关键认知失真信息 3.通过多特征自适应加权方法为不同的用户动态识别关键的后期融合信息和多个用户级别的信息
O	输出	用户的状态是抑郁还是正常

P	问题	1. 忽略了关于用户认知扭曲信息 2.只关注用户帖子的语义信息，而忽略了与抑郁相关的各种用户层面信息
C	条件	用户级、多帖子、带个人属性的数据
D	难点	1.抑郁不是单帖特征，而是长期用户状态 2.社交媒体文本非常嘈杂，用户级行为特征不是每个人都一致出现 3.同一场景中不同认知扭曲要区分
L	水平	2026 SCI中科院1区

2DM-CDUI 算法原理

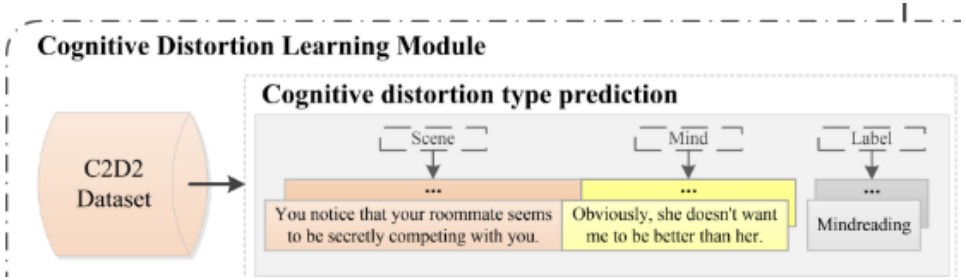
- 2DM-CDUI由三个模块组成：
  - 认知扭曲学习模块：捕捉认知扭曲的编码器MSCDE
  - 帖子级信息融合模块：同时提取帖子语义表示和认知扭曲表示，并通过“帖子级认知扭曲感知机制”突出与用户主观认知相关的扭曲信息
  - 用户级信息融合模块：融合帖子级表示 + 性别 + 发帖行为特征，并通过“多特征自适应加权方法”动态决定不同信息的重要性



2DM-CDUI 认知扭曲学习模块

- C2D2数据集使用
  - 借助C2D2数据集学习认知扭曲知识。C2D2里包含了“scene（场景）”和“mind（对应想法）”以及认知扭曲类型标注。八种典型认知扭曲类别，7500条高质量文本，涉及300个负面事件与场景
- 认知扭曲类型预测
  - 将C2D2数据集中每个数据项中的场景文本和对应的想法文本拼接在一起，并用作输入文本
  - 使用编码器进行编码
  - 训练获得文本的上下文语义表示
  - 输入到Softmax层，以获得文本的认知扭曲类型

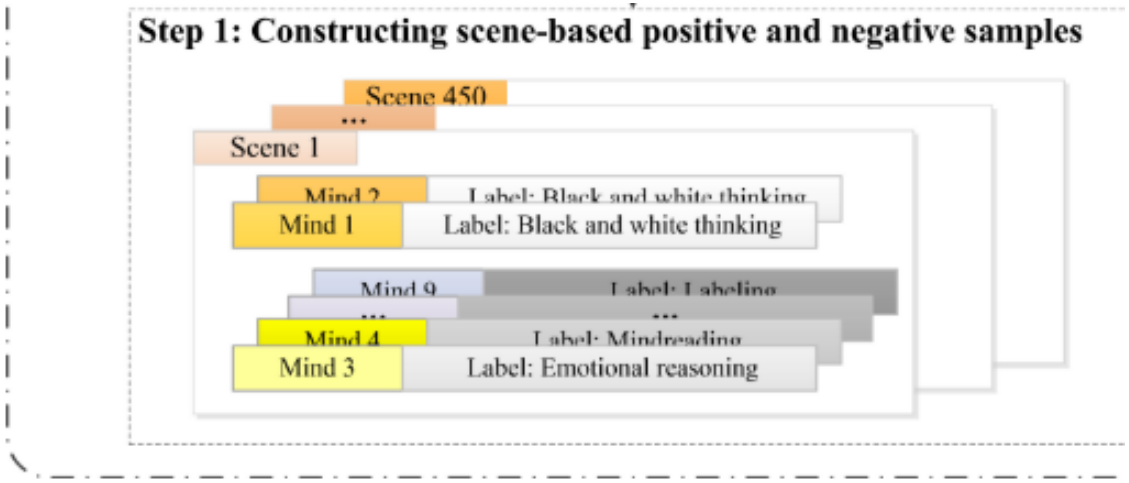
Num	场景	思维	标签
1	很内向的人，你刚到了一个新环境，周围的人都比较陌生。	这个环境里的人是不是都很不友善啊	过度泛化
2	最近你感觉有时候站起来就头晕，感觉头重脚轻那种	头好晕，是不是生病了，得去医院看看了	非扭曲
3	走在路上你感觉饿了，但周围一家饭店都没有	累了没地方休息，饿了没地方吃饭，怎么不幸的事情总是发生在我的身上呢	过度泛化



让模型学到粗粒度的认知扭曲类别边界

2DM-CDUI 认知扭曲学习模块

- 场景对比学习预训练
  - 构造基于场景的正负样本：按场景分类数据，构建一组同一场景类型的想法文本
  - 在同一场景内部，再按认知扭曲类别分组
  - 构造训练样本
    - 先选一个类别作为“正类”：两个样本，一个作为学习样本h^l，一个作为正样本h^{po}
    - 从其他类别里各取一个作为负样本h^{ne}，构成一个正负样本组
  - 探索在相同场景下，不同扭曲的表达边界在哪里



2DM-CDUI 认知扭曲学习模块

- 场景对比学习预训练
  - 把所有文本依次输入编码器

$$lh^{cl} = \frac{\exp(sim(h^l, h^{po})/\tau)}{\exp\left(\frac{sim(h^l, h^{po})}{\tau}\right) + \sum_{i=0}^{7} sim(h^l, h^{ne})}$$

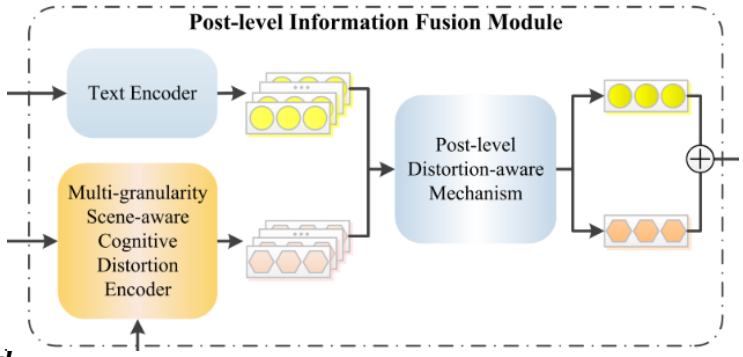
$$Loss^{cl} = \sum_{i=1}^{k} -log lh_i^{cl}$$

  - 同场景、同扭曲类型 → 拉近
  - 同场景、不同扭曲类型 → 拉远

让模型从类别级知识进一步提升到场景感知的细粒度认知表达知识，最终得到多粒度场景感知认知扭曲编码器MSCDE

2DM-CDUI 帖子级信息融合模块

- 双路编码
  - 对于用户帖子集合P中每一条帖子：
    - 用普通预训练语言模型 PLM 编码得到上下文语义表示H^{cs}
    - 用MSCDE编码得到认知扭曲表示H^{cd}
- 映射到同一空间
  - 通过同一个全连接层做线性映射到同一维度，得到$\hat{H}^{cs}$和$\hat{H}^{cd}$
- 计算用户全局语义
  - 把帖子集合的语义表示做类似self-attention的计算，得到一个全局语义表示H^{csr}



$$L_{mean} = [\frac{1}{n}, \dots, \frac{1}{n}]$$

$$H^{csr} = \sigma(\frac{L_{mean}\hat{H}^{cs}W_1^{pcd}(\hat{H}^{cs}W_1^{pcd})^T}{\sqrt{d_p}})$$

把所有帖子聚合成一个能代表该用户总体表达语义的向量

2DM-CDUI 帖子级信息融合模块

- 计算每个帖子对应的认知扭曲综合表示
  - 对每条帖子的认知扭曲表示也做综合处理，得到H^{ccd}

$$H^{ccd} = \{h_1^{ccd}, \dots, h_n^{ccd}\}$$

让每条帖子的认知扭曲信息不只停留在独立向量，而是也带有“全局上下文”感知

- 帖子级认知扭曲感知机制
  - 把某条帖子的认知扭曲表示和用户整体语义表示做乘积，再经过Softmax得到每个帖子的权重Z^r

$$H^r = \{h_1^{ccd} \odot H^{csr}, \dots, h_n^{ccd} \odot H^{csr}\}$$

$$Z^r = Softmax(H^r W_5^{pcd})$$

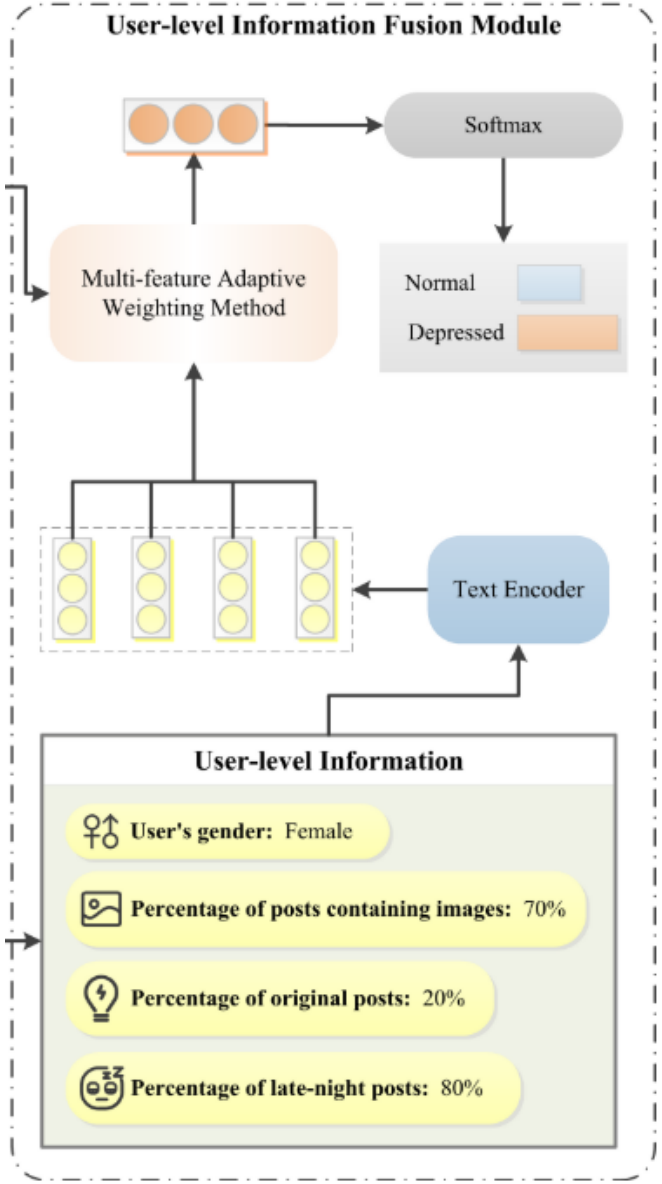
通过“全局语义×帖子认知扭曲”的关联度，筛出更可信的扭曲信息

  - 权重越高，说明该帖子中的认知扭曲越符合该用户整体语义和主观认知，将计算得到的每个帖子的重要性权重与映射的认知失真表示相乘，得到增强的认知失真表示H^{ecd}

$$H^{ecd} = \{h_1^{ecd}, \dots, h_n^{ecd}\} = \{Z_1^r \cdot \hat{h}_1^{cd}, \dots, Z_n^r \cdot \hat{h}_n^{cd}\}$$

2DM-CDUI 用户级信息融合模块

- 使用四类用户级信息：
  - 性别、原创帖比例、含图片帖比例、深夜发帖比例（23:00 到次日 6:00）
- 将用户信息文本送入PLM
  - 把异构特征变成文本信息，与帖子文本表示在同一种语义空间中融合
- 多特征自适应加权方法
  - 拼接成 5 个特征向量组成的集合H^u，然后通过一个类似self-attention的方式计算每个特征的权重α_i
- 将加权后的用户级特征做平均池化，得到最终用户表示，再送到softmax做分类



实验设计　数据资源

- 数据集
  - SWDD数据集：按时间对用户帖子重新排序，再清洗噪声数据。作者指出原始SWDD存在明显类别不平衡，因此构建了两个平衡子集：
    - SWDD-T1：3600个样本
    - SWDD-T2：7400个样本
  - TMDD数据集
    - 在Twitter Mental Disorder Dataset (TMDD)上实验。由于该数据集缺少用户性别、发帖时间等信息，因此这里只能用去掉用户行为信息的消融版本进行比较
- 评价指标
  - Accuracy
  - Macro-F1
  - Precision
  - Recall

Dataset		Total	Depressed	Normal
SWDD-T1	Train	2880	1440	1440
	Test	720	360	360
SWDD-T2	Train	5920	2960	2960
	Test	1480	740	740
TMDD	Train	13554	6777	6777
	Test	3388	1694	1694

实验设计　对比实验

- 传统模型对比
  - 在SWDD-T1和SWDD-T2上整体优于各类 PLM 基线
- 引入认知扭曲和用户级信息后，在不同数据规模下都增益
- 传统模型大多只关注帖子语义，即便Chinese MentalBERT引入精神健康知识，也没有同时建模：认知扭曲性别用户行为因此表现不如2DM-CDUI

Models		SWDD-T1				SWDD-T2			
		Acc.	F1	Prec.	Rec.	Acc.	F1	Prec.	Rec.
PLMs	BERT (2018)	85.42	85.41	84.66	86.17	84.79	84.79	84.74	84.85
	BiLSTM (2020)	87.63	87.63	88.21	87.05	87.83	87.82	87.20	88.48
	BiLSTM+Att (2022)	86.25	86.24	85.90	86.60	87.36	87.36	86.49	88.25
	HCN+ (2023)	86.94	86.91	87.02	86.79	88.17	88.17	88.00	88.34
	CBA (2024)	87.77	87.70	88.00	87.41	87.50	87.48	86.98	88.00
	Chinese MentalBERT (2024)	88.33	88.31	88.00	88.62	88.85	88.85	87.72	90.00
Ours	2DM-CDUI	**90.99**	**91.11**	**90.50**	**91.70**	**90.47**	**90.25**	**90.00**	**90.40**

实验设计　对比实验

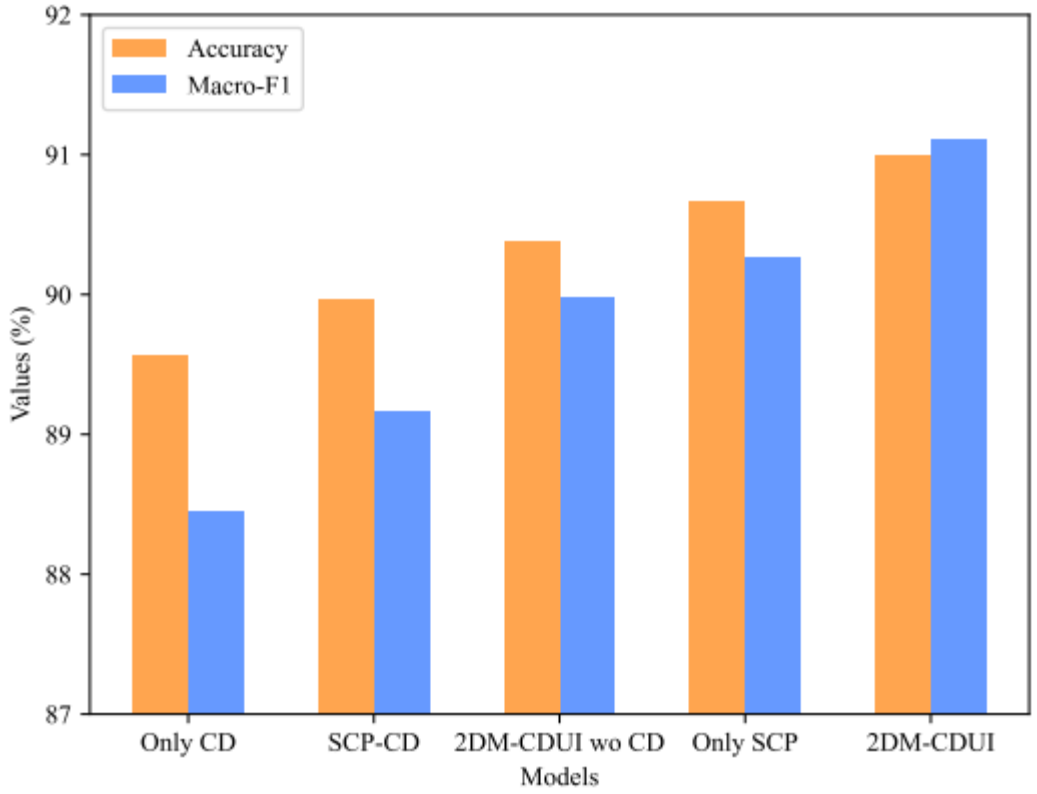
- 大模型对比
  - GPT-4o、DeepSeek、Llama3+LoRA在部分结果上更强
  - 但2DM-CDUI参数量只有103M，却能接近甚至超过一些参数量远大得多的 LLM
- 2DM-CDUI的优势不是绝对顶尖性能，而是参数更小训练成本更低

	Models	Params	SWDD-T1				SWDD-T2			
			Acc.	F1	Prec.	Rec.	Acc.	F1	Prec.	Rec.
LLMs	GPT-4o mini (2024)	8B	72.00	70.58	81.33	73.58	77.00	76.00	83.09	79.09
	Llama3 (2024)	70B	86.00	84.49	82.00	87.06	89.00	89.00	87.00	91.00
	Llama3+LoRA (2024)	8B	90.00	89.85	90.78	89.60	93.00	**92.94**	92.87	**93.03**
	Mixtral of Experts (2024)	56B	85.00	85.00	85.47	85.37	90.00	89.98	90.12	90.51
	GPT-4o (2024)	200B	**92.00**	**91.97**	**91.97**	**91.97**	**93.00**	92.90	**93.02**	92.83
	MentalLLAMA (2023)	13B	81.00	80.66	81.36	80.01	84.00	83.78	83.50	84.06
	DeepSeek (2025)	671B	91.00	91.00	91.51	91.39	91.00	90.96	90.87	91.21
Ours	2DM-CDUI	103M	90.99	91.11	90.50	91.70	90.47	90.25	90.00	90.40

实验设计　消融实验

- 评估各个模块的作用
  - w/o PDM：去掉帖子级扭曲感知机制
  - w/o UPBI：去掉用户画像和行为信息
  - w/o MAWM：去掉多特征自适应加权
  - w/o CD：去掉认知扭曲学习模块

Models	SWDD-T1			
	Acc.	F1	Prec.	Rec.
2DM-CDUI w/o PDM	88.94	89.30	88.86	89.74
2DM-CDUI w/o UPBI	89.86	89.56	88.66	90.48
2DM-CDUI w/o MAWM	90.24	89.89	89.54	90.24
2DM-CDUI w/o CD	90.38	89.98	90.11	89.85
2DM-CDUI	**90.99**	**91.11**	**90.50**	**91.70**



算法总结　2DM-CDUI

- 算法贡献
  - 首次将认知扭曲引入用户级抑郁检测，使模型不仅看“发了什么”，还看“用户以何种扭曲认知方式表达”
  - 提出多粒度认知扭曲学习方法，通过“扭曲类型预测 + 场景对比预训练”同时学习粗粒度类别信息和细粒度场景差异
  - 设计了帖子级扭曲感知和用户级自适应加权融合机制，能够动态筛选更关键的帖子信息与用户行为信息

- 算法不足
  - 模型依赖外部认知扭曲数据集和人工构造的用户特征，跨平台或跨领域泛化能力仍有不确定性
  - 实验主要基于重采样后的平衡数据集，和真实环境存在差距

特点总结与未来展望

特点总结与工作展望

- 特点总结
  - ELB
    - 将认知扭曲检测提升为实例级推理+bag级聚合的MIL框架
    - 提出ELB心理结构分解，将一句话拆成情绪、逻辑、行为三个维度，为后续推理提供更明确的心理学支架
  - 2DM-CDUI
    - 首次将认知扭曲系统性引入用户级抑郁检测任务，不再只依赖帖子表层语义，而是利用更接近抑郁心理机制的认知偏差特征
    - 提出帖子级扭曲感知机制和用户级自适应加权融合机制，从多条帖子中筛选与用户整体主观认知更相关的关键扭曲信息，减少转发、噪声文本的干扰
- 工作展望
  - 从单一文本模态走向用户行为、对话上下文、多模态信息融合
  - 从静态文本判断走向动态时序建模

参考文献

[1] Kim J S, Kim H, Oh W J, et al. Multi-View Attention Multiple-Instance Learning Enhanced by LLM Reasoning for Cognitive Distortion Detection[J]. arXiv preprint arXiv:2509.17292, 2025.

[2] Wan Y, Dong Z, Jiang B, et al. Enhancing Depression Detection with Cognitive Distortion and User-level Information[J]. IEEE Journal of Biomedical and Health Informatics, 2026.

知人者智，自知者明。胜人者有力，自胜者强。知足者富。强行者有志。不失其所者久。死而不亡者，寿。

谢谢！