信息网络专题研究之应用层-机器学习

一、文献信息

1. 论文题目：Assessing Viewer’s Mental Health by Detecting Depression in YouTube Videos

2. 作者：Shanya Sharma SAP Labs，andManan Dey SAP Labs

3. 发表途径：NeurIPS Joint Workshop on AI for Social Good

4. 发表时间：December,2019

二、问题意义

1. 研究背景和意义

抑郁症是世界上最普遍的心理健康问题之一，也是自杀的主要诱因。抑郁症是世界第四大疾病，而到2020年它上升为第二大疾病。然而大众对抑郁症的医疗防治还处在诊断率低的局面，只有五分之一的患者主动求医，十分之一的患者接受了相关的药物治疗；与此同时抑郁症的发病（和自杀事件）已开始出现低龄（大学，乃至中小学生群体）化趋势。因此，对抑郁症的科普、防范、治疗工作亟待重视。

按照临床标准，一个人持续两周陷入低落情绪可能被诊断为抑郁症。

研究显示，精神抑郁的人往往会反复观看含有消极情绪的视频(抑郁视频)，以便获得认可或找到同伴。这种行为模式在确定用户的心理状态时非常有用，因此可以通过分析一段时间内的观看历史预估用户的精神状态。

2. 研究问题

(1) 如何利用深度学习分辨抑郁视频与非抑郁视频

(2) 判定结果与实际结果的吻合情况

(3) 提出了一种记录并评估用户观看视频类型来预测用户心理状态的方案，用以发现有抑郁症倾向或患有轻度抑郁症的用户。

三、思路方法

本文对抑郁视频的判定是基于文本特征对视频内容进行分析，构建分类器来识别视频是否属于抑郁。作者通过关键字将视频分为四类：a）音乐 b）抑郁 c）有趣 d）自助/激励。从每个类型中随机选取约200个视频，分析视频的抑郁程度，并通过视频评论的CES-D分数验证。

1. 三种分析模型

三种模型包括：1）移情模型+基于 N-Grams 模型的朴素贝叶斯算法（NB withN-Grams）；2）词频-逆文本频率指数（TF-IDF）+移情模型+基于 N-Gram 模型的朴素贝叶斯算法 3）长短期记忆网络（LSTM）。

- 移情模型（Empath model）

移情模型是一种结合现代自然语言处理技术对文本词汇中包含的情绪进行量化分析的模型。除了预先定义的积极和消极情绪等特征外，移情模型还可以创建自定义特征。根据抑郁症现象对应的关键词，如厌食、情绪低落、失眠、自我厌弃、难以集中注意力等构建类别词汇。例如，对于关键词“减肥”，生成的一组词是“减肥”、“贫血”、“压力”、“营养不良”。添加自定义关键词使文本的分析更加准确。

- 基于 N-Grams 模型的朴素贝叶斯算法（Naive Bayes with n-grams:）

N-Grams 模型基于马尔科夫假设，第 n 个词的出现只与前面 N-1 个词相关，而与其它任何词都不相关，整句的概率就是各个词出现概率的乘积。这些概率可以通过直接从语料中统计 N 个词同时出现的次数得到。概率越高，是整句（合理表达）的可能性就更大。而朴素贝叶斯算法是对特征值进行独立性假设，这样每个特征值对结果的影响互不相关，可以有效减少计算量，提高计算速度。

- 词频-逆文本频率指数（TF-IDF）

字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降。TFI-DF 的主要思想是：如果某个词或短语在一篇文章中出现的频率高，并且在其他文章中很少出现，则认为此词或者短语具有很好的类别区分能力，适合用来分类。在模型中，它可以用于过滤高频率但情感中性的词语，提高分类准确性。

- 长短期记忆网络（LSTM）

LSTM 是一种特殊的循环神经网络，通过门控状态来控制传输状态，记住需要长时间记忆的，忘记不重要的信息；而不像普通的 RNN 只能不分轻重地叠加记忆。嵌入层的作用主要在于学习词语的分布式表达，将词句构成的稀疏向量降维成包含更多语义和语法信息的密集向量。将嵌入层的输出送入多单元（196 单元）、用 tanh 激活的的 LSTM 网络中递归运算，输出视频分类结果。

2. 评论的评估机制——CES-D 分数

为了评估分类器的准确程度，作者对视频进行 CES-D *评分*。当出现在 CES-D 量表中的症状以负面形式出现在评论中（例如对应食欲出现厌食，对应睡眠出现失眠）时，视频的 CES-D 分数将提高。具体算法是，将评论涉及的关键词并入一个集合，通过遍历集合中的词

计算词频，并通过累计词频计算聚合。由于集合中的词可以同时用于积极和消极的含义，因此将评论发送到移情模型，以分析积极和消极情绪的得分，如果积极情绪>消极情绪，则CES-D 分为 0 分。将所有评论的 CES-D 分数平均，生成标准化分数。

四、实验结论

结合评论的 CES-D 评分，三种分析模型的准确度不尽相同。模型 1(EMPATH+NB)的准确度只有 52%，但加入 TF-IDF 的模型 2（TF-IDF+EMPATH+NB）准确度为 81.2%，模型 3（LSTM）准确度为 83.4%。从中可以推断，移情模型和 TF-IDF 的结合可以产生了几乎与 LSTM 相同的分类精度，且由于自身 LSTM 网络的局限，模型 3 所需时间（345s）远高于模型 2（8s）。

将视频按情节内容分为音乐、幽默、抑郁、自我激励四类后，作者发现在音乐、幽默的视频中非抑郁评论百分比（CESD=0）远高于抑郁类别。且在抑郁视频的评论中 CES-D 评分是最高的，这证明抑郁是由抑郁诱发和吸引的。由此可以推断，频繁观看 CES-D 分数高的视频（抑郁视频）很可能是一种不健康的精神状态的指标。

在对为训练目的而收集的抑郁视频评论区进行 CES-D 评分平均后，作者得出将阈值设置为 20 的结论。为了评估分类器在真实环境中产生的分类结果，作者随机挑选 1500 个视频输入模型，并将分类结果与评论所得的 CES-D 评分进行比较，可得出基于上述方法设置分类器的准确度为 84%，

五、启发思考

本文提出了一种分析视频所蕴含情绪的方法，通过分析视频的文本特征，判断视频的类型，并结合用户在一段时间内的观看历史，判断用户是否有不健康心理状态的迹象，期望在用户陷入抑郁情绪的初期就将其引导到正确的门户网站寻求帮助。

在国内常用的社交软件、视音频网站和搜索引擎中（如新浪微博、网易云音乐、哔哩哔哩弹幕网、百度搜索）中，搜索“自杀”相关词句，会出现全国 24 小时免费心理咨询电话以及其他心理干预。与这种目前现有的、被动的自杀干预机制相比，本文提出的基于观看历史的心理分析预警更偏向于早期发现、主动提醒，将是一种行之有效的干预机制。

本文分析视频的性质主要是依靠视频的文本特征，但我们在观看视频时常常感受到“沉默的表达”更为深刻，一些具有视觉冲击性的、震撼的画面更容易引起情感共鸣。是否可以通过文本检测结合边缘检测、图像分割等图像识别方法，以及语音识别对视频的情绪进行分析，提高视频区分的准确性。

此外，利用机器学习对视频的分析不仅可以用于抑郁症的早期发现，在建立完整的视频库、语料库后，也可用于有害信息筛查、不良信息过滤，以及更广泛的视频分级制度。