

016 | 精读2017年EMNLP最佳长论文之二

2017-11-08 洪亮劼

AI技术内参

[进入课程 >](#)



讲述：初明明

时长 08:55 大小 4.09M



EMNLP 每年都会选出两篇最佳长论文，我们已经分析过第一篇《男性也喜欢购物：使用语料库级别的约束条件减少性别偏见的放大程度》。今天我继续来讲第二篇。

EMNLP 2017 年最佳长论文的第二篇是《在线论坛中抑郁与自残行为风险评估》

(Depression and Self-Harm Risk Assessment in Online Forums)。这篇文章探讨了利用自然语言处理技术来解决一个社会问题。最近一段时间以来，如何利用机器学习、数据科学等技术来解决和处理社会问题，正逐渐成为很多社会科学和机器学习研究的交叉领域。

作者群信息介绍

第一作者安德鲁·耶特斯 (Andrew Yates)，计算机博士，毕业于美国华盛顿的乔治城大学 (Georgetown University)，目前在德国马克斯普朗克信息学院 (Max Planck Institute

for Informatics) 攻读博士后。他在博士阶段已经发表了多篇采用深度学习技术和信息检索、自然语言处理相关的论文。

第二作者阿曼·可汗 (Arman Cohan)，来自伊朗，是乔治城大学计算机系博士生。阿曼已在信息检索和自然语言处理相关方向发表了多篇论文。2016 年，在华盛顿的 Medstar Health 实习并发表了两篇论文。2017 年暑假，在美国加州圣何塞 (San Jose) 的奥多比 (Adobe) 研究院实习。

第三作者纳兹利·哥汗 (Nazli Goharian) 也来自乔治城大学计算机系，目前在系里担任计算机教授。第一作者是他之前的学生，第二作者是他当前的学生。纳兹利在长达 20 年的职业生涯中先后在工业界和学术圈任职，可以说有很深厚的学术和工业背景，他在信息检索和文本分析领域已发表 20 多篇论文。

论文的主要贡献

在理解这篇文章的主要贡献之前，我们还是先来弄明白，这篇文章主要解决了一个什么场景下的问题。

现代社会，人们生活工作的压力越来越大。研究表明，很多人都可能受到各式各样精神疾病 (Mental Conditions) 的困扰。在当下发达的互联网时代，在线场所为这些精神疾病患者寻求帮助提供了大量的资源和信息，特别是一些专业的在线支持社区，或是一些更大的在线社区比如 Twitter 或者 Reddit。

因此，研究这些人在各种在线社区的行为，对设计更加符合他们需要的系统有很大帮助。对于很多社会研究人员来说，分析这些人的精神状态，才能更好地帮助他们长期发展。

这篇文章提出了一个比较通用的框架，来分析这些精神疾患者的在线行为。在这个框架下，可以比较准确地分析发布信息的人是否有自残 (Self-Harm) 行为，还可以比较容易地分析哪些用户有可能有抑郁症 (Depression) 的状况。

整个框架利用了近年来逐渐成熟的深度学习技术对文本进行分析。所以，这里的应用思路很值得借鉴和参考，也可以用于其他场景。

论文的核心方法

在介绍这篇文章提出的方法之前，作者们用不小的篇幅介绍了文章使用的数据集和如何产生数据的标签。

首先，作者们从著名的在线社区 Reddit 中找到和精神疾病有明确联系的帖子。这些帖子是按照一个事先准备的语料库来筛选的，这个语料库是为了比较高精度地发现与精神疾病相关的帖子。利用语料库里的句式，比如“我已经被诊断得了抑郁症”，这样就可以保证，找到的帖子在很大程度上是来自精神疾病患者的。

如果一个用户发布了这样的帖子，但在这之前发布的帖子少于 100 条，这个用户就不会包含在数据库中。做这样的筛选可能作者们的考虑是，太少的帖子无法比较全面地包含用户方方面面的行为。

作者们在 Reddit 社区中挖掘了从 2006 年到 2016 年十年时间里符合条件的所有帖子，并利用人工标注的方式筛选出了 9210 个有精神疾病困扰的用户。这些可以当做机器学习的正例。

那么如何寻找负例呢？作者们当然可以利用所有的用户，但是这样带来的后果很可能是研究没有可比性。如果正例的用户和负例的用户之间差别太大，我们就很难说这些差别是因为精神疾病造成的还是由其他区别带来的。于是，作者们想到的方法则是尽可能地对于每一个正例的用户都找到最接近的负例用户。

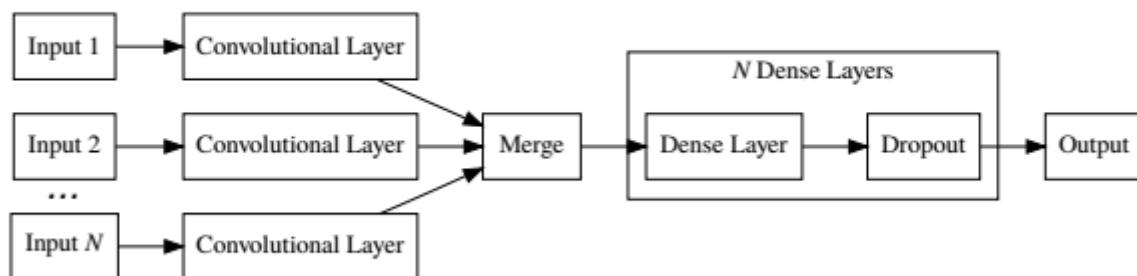
实际操作中，作者们采取了更加严格的方式，那就是负例的用户必须没有发布过任何与精神疾病相关的帖子，并且在其他方面都需要和正例用户类似。在这样的条件下，作者们找到了 107274 个负例用户。

对于数据集中的用户而言，每个用户平均发布 969 个帖子，平均长度都多于 140 个字。可以说，由这些用户构成的这个数据集也是本文的一个主要贡献，这个数据集用于分析抑郁症。

对于自残行为而言，作者们利用了一个叫 ReachOut 的在线社区的数据，收集了包括 65024 个论坛的帖子，其中有 1227 个帖子提到了自残。而对于提及自残的程度，数据分了五个等级用于表示不同的紧急情况。

这篇论文主要提出了基于卷积神经网络的文本分析框架，分别用于检测抑郁症用户和检测自残倾向度的两个任务中。虽然这两个任务使用的数据不同，最终采用的模型细节不同，但是

两个任务使用的都是同一个框架。下面我就来说一说这个框架的主要思想。



首先，作者们利用每个用户的发帖信息来对每一个用户进行建模，基本的思路是通过神经网络来对用户的每一个帖子建模，从中提取出有效信息，然后把有效信息汇总成用户的一个表达。有了这个思路，我们再来看看具体是怎么做的。

每个帖子一个范围内的单词首先通过卷积层（Convolutional Layer）提取特征，然后提取的特征再经过最大抽取层（Max Pooling Layer）集中。这个步骤基本上就是把目前图像处理的标准卷积层应用到文本信息上。每一个帖子经过这样的变换就成了特征向量（Feature Vector）。有了这样的特征向量之后，用户的多个特征向量整合到一起，根据不同的任务形成用户的整体表征。

在检测抑郁症的任务上，作者们采用的是“平均”的方式，也就是把左右的帖子特征向量直接平均得到。而在检测自残的任务上，作者们则采用了一种比较复杂的形式，把所有的帖子都平铺到一起，然后再把当前帖子之前的帖子，作为负例放在一起，注意，不是平均的形式，而是完全平铺到一起，从而表达为用户的整体特征。

在经过了这样的信息提取之后，后面的步骤就是构建分类器。这个步骤其实也是深度学习实践中比较常见的做法，那就是利用多层全联通层（Fully Connected Layer），最终把转换的信息转换到目标的标签上去。

可以说在整体的思路，作者们提出的方法清晰明了。这里也为我们提供了一种用深度学习模型做文本挖掘的基本模式，那就是用卷积网络提取特征，然后通过联通层学习分类器。

方法的实验效果

作者们上面提到的实验数据集上做了很充分的实验，当然也对比了不少基本的方法，比如直接采用文本特征然后用支持向量机来做分类器。

在辨别抑郁症的任务上，本文提出的方法综合获取了 0.51 的 F1 值，其中召回 (Recall) 达到 0.45，而直接采用支持向量机的方法，精度 (Precision) 高达 0.72，但是召回指数非常低只有 0.29。

而在检测自残的任务上，提出方法的准确度能够达到 0.89，F1 值达到 0.61，都远远高于其他方法。

应该说，从可观的数值上，本文的方法效果不错。

小结

今天我为你讲了 EMNLP 2017 年的第二篇年度最佳长论文，这篇文章介绍了一个采用深度学习模型对论坛文本信息进行分析的应用，那就是如何识别有精神疾病的用户的信息。

一起来回顾下要点：第一，我简要介绍了这篇文章的作者群信息。第二，这篇文章是利用自然语言处理技术解决一个社会问题的应用，论文构建的数据集很有价值。第三，文章把目前图像处理的标准卷积层应用到文本信息上，提出了基于卷积神经网络的文本分析框架，用于辨别抑郁症和检测自残倾向，都实现了不错的效果。

最后，给你留一个思考题，如果说在图像信息上采用卷积层是有意义的，那为什么同样的操作对于文本信息也是有效的呢？文本上的卷积操作又有什么物理含义呢？

欢迎你给我留言，和我一起讨论。

拓展阅读：[Depression and Self-Harm Risk Assessment in Online Forums](#)


AI 技术内参

你的360度人工智能信息助理

洪亮劼

Etsy 数据科学主管
前雅虎研究院资深科学家



新版升级：点击「 请朋友读」，10位好友免费读，邀请订阅更有**现金**奖励。

© 版权归极客邦科技所有，未经许可不得传播售卖。页面已增加防盗追踪，如有侵权极客邦将依法追究其法律责任。

上一篇 015 | 精读2017年EMNLP最佳长论文之一

下一篇 017 | 精读2017年EMNLP最佳短论文

精选留言 (4)

 写留言



徐听听

2017-11-08

 6

图像上用卷积是利用了卷积作用在图像像素的空间平移不变性，例如一只猫无论是出现在图像的哪个位置（左上还是右下角），都对分类器结果无影响。

将卷积作用在文本信息中能够有效应该也是利用了卷积的平移不变性，无论和抑郁/自残相关的词句出现在文本的哪个位置，只要出现过，都可以被检测出来。

展开 



阿卡斯

2018-07-06

 1

对于那个模型我有两个技术细节不是特别清晰，，最开始输入的是文本（即帖子文章）我

可以理解成是应用了word2vec向量化文章中每一个单词构成了大矩阵进入到CNN么？那这样的话每个文章字数是不同的，我们即使确定了文章最大长度的95%还是会有信息缺失，这个输入矩阵宽度如何设计？第二个问题每个用户的帖子文章数量是不同的最后我们通过CNN提取feature数量是不同的具体是怎么merge没有提到

展开 ∨



huan

2017-11-09

👍 1

我的理解是，这里的场景下，物理意义是帖子发表人表达抑郁或者自残情绪的时候，选词是比较狭窄的，负面的，和正常人比较是相对偏少的，而且抑郁或者自残的情绪波动比较少（正常人的情绪应该比较多）。不过不知道这种“物理意义”是否存在很多的偏见。另外不知道文章中的原始输入X是怎么做向量化的，也就是“每个帖子的一个范围内单词”是怎么操作的？直接分词还是做stem，或者去掉噪音词吗？保留词的顺序吗？没有...

展开 ∨

作者回复: 我的理解是用了Embedding。



徐听听

2017-11-08

👍

图像上用卷积是利用了卷积作用在图像像素的空间平移不变性，例如一只猫无论是出现在图像的哪个位置（左上还是右下角），都对分类器结果无影响。

将卷积作用在文本信息中能够有效应该也是利用了卷积的平移不变性，无论和抑郁/自残相关的词句出现在文本的哪个位置，只要出现过，都可以被检测出来。

展开 ∨