笔记 4

1 文献信息

作者：Anqi Liu, Maya Srikanth, Nicholas Adams-Cohen, R. Michael Alvarez, Anima Anandkumar

论文题目：Finding Social Media Trolls: Dynamic Keyword Selection Methods for Rapidly-Evolving Online Debates

发表途径：AI for Social Good Workshop NeurIPS2019

发表时间：2019 年 12 月

2 问题意义

2.1 研究问题

网络骚扰是当今一个重大的社会问题。开发较少偏见、更快且更有效的方法来识别社交媒体的负面情绪，仇恨言论和在线骚扰非常重要。根据皮尤研究中心（Pew Research Center）2017 年的调查数据，有 41％的美国人自己经历过网上骚扰，而 66％的美国人表示他们看到过针对他人的网上骚扰。在这些自己经历过网上骚扰的美国人中，有 18％的人表示自己受到了严重的骚扰（例如，长期受到身体威胁或性骚扰等）。

要想防止网络骚扰，就需要迅速检测到那些制造骚扰的负面的社交媒体贴文。这些数据大部分是通过关键字搜索来收集的。但是，以往的关键字检测方法并不适合动态的在线辩论。在本文中，作者提出了一种针对动态辩论进行关键字检测的模型，即 GloVe（Global Vectors for Word Representation, 全球单词表示向量）模型，并建议使用单词嵌入模型从两个方面识别攻击性和骚扰性社交媒体消息：检测快速变化的主题以更有效地收集数据，以及在不同领域表达单词语义。

2.2 研究背景

社交媒体数据，特别是来自 Twitter 的数据，已被用于许多社会、政治和经济行为的研究中。这些数据大部分是通过关键字搜索来收集的，例如收集包含特定关键字或主题标签的推文。使用 Twitter 上的特定关键字或标签来收集社交媒体数据的方法非常容易，并且为许多研究人员所使用。过去的工作中一般有三种方法来搜索社交媒体网站以获取与研究相关的材料：

（1）完全非自动方式

研究人员根据他们对主题的了解，事先设定一组关键字。这种方法易于实现且相当透明，

但它是静态的，很容易导致数据有偏差。如果要对一个长时期内的对话进行检测，那么研究人员将需要以完全人工的方式更新静态关键字列表。

（2）完全自动方式

这种方法是可以自动更新搜索关键字的自动关键字选择方法，例如电子邮件垃圾邮件过滤器所使用的方法。这些通常基于非透明的机器学习或深度学习方法，研究人员很难对其进行评估。此外，数据挖掘社区提出的全自动关键词提取方法主要集中在减少人工标注工作，提高准确性和增大静态主题的覆盖范围上。但是，要对涉及动态主题的在线辩论进行建模和跟踪，就需要检测语言和措辞的演变，以及快速适应学习算法的新数据域。

（3）半自动方式

这种方法是半自动关键词选择方法，它基于人和机器的分析相结合来确定合适的关键词，这种方法同时具有完全自动和完全非自动关键字选择方法的优缺点。

2.3 研究意义

研究 GloVe（Global Vectors for Word Representation）模型的意义在于，以往使用的方法不适合动态的在线辩论，因为动态对话中的主题标签和关键字会随着时间发生快速变化。社交媒体用来辩论特定问题的术语在较长的时间范围内也会发生很大变化，并且随着事件发展，人们会改变对问题的讨论方式，社区可能会使用不同的标签和关键字。当涉及监视和研究消极情绪，如仇恨言论以及其他形式的负面社交媒体行为（如诈骗和网络骚扰）时尤其如此。这些类型的在线对话可能是高度动态的，从事这些讨论的人一般会经常更改关键字和主题标签，以避免社交媒体平台本身对其进行检测并采取行动。使用 GloVe 模型可帮助发现新的关键字以用于数据收集和拖钓检测。

3 思路方法

3.1 研究思路

文章首先讨论了过去有关关键字搜索的研究，随后提出了一种基于 GloVe 模型的动态关键字搜索的新方法，并将其用于来自#MeToo 运动（一个反对性骚扰的社交媒体运动）的数据的初步分析。在对数据进行分析后，作者发现利用单词向量进行关键字检测有可能发现有更意义的关键字，这些关键字可能会暴露以前看不见的线上交互。在结尾部分，作者阐述了动态关键字搜索方法的开发和使用的后续步骤。

3.2 研究方法

作者的目标是开发一种快速有效的动态关键字搜索和更新方法，这种方法可提供透明性

并减少偏差。在迅速发展的社交媒体讨论中，非自动化的关键字选择效率低下，成本高昂且耗时。对于半自动方式来说，尽管它可以通过人工干预来改善关键字选择，但是在涉及负面社交媒体的讨论中，人工参与关键字选择处于道德上的考虑应避免使用（例如，使用人工标记具有攻击性的社交媒体帖子）。

本文提出一种新的检测方法，它建议使用单词嵌入模型从一组静态关键字收集的数据中学习单词的表示形式。文章提供了两种方法来提取关键字：（1）提取最频繁出现的单词；（2）使用聚类算法来发现“簇”。对于每个簇，使用余弦相似度或排序算法选择关键字。使用这些关键字作为数据并重新训练模型后，再以类似的方式获取下一组查询。初步结果证明，结合单词嵌入模型有助于跟踪动态主题的演变。这项工作为将来的拖钓检测铺平了道路。

当利用词嵌入模型进行关键词提取时，作者观察到了不同讨论社区之间的明显差异。Red Pill subreddit 社区是一个男性权利社区，也是#MeToo 运动的反对者。截至 2019 年 3 月，在 Red Pill subreddit 上经过训练的单词嵌入，与包含#MeToo 主题标签的 Twitter 上的经过训练的单词嵌入相比存在相当的差异。同时，作者还纳入了在 Wikipedia 2014 和 Gigaword5 语料库中经过预训练的单词嵌入作为比较。在未来，在这些社区中使用经过训练的单词嵌入，将有可能揭露更多其他的负面和仇恨言论。

本文讨论过的方法，已经经过实验测试。下一步的工作包括进一步开发、更好地实现自动化以及使用功能更强大的 Twitter API（允许访问更大的推文数据库）。作者还计划以几种不同的方式测试和验证他们的方法，比如将他们的方法与人类分析人员进行比较，以评估该方法相对于人工检测的速度和准确性。作者还将该方法与其他类似的动态关键字检测方法进行比较，比较的指标依然是速度和准确性。

4 实验结论

本文提出一种新的动态关键字搜索和更新方法，这种方法可提供检测的透明性并减少误差。初步结果证明，结合词嵌入模型有助于跟踪动态主题的演变。这项工作为将来的拖钓检测铺平了道路。这种方法可以更好地理解通过社交网络传播的信息和情感，并为最终通过社交网络描述信息和情绪的变化，提出一种机制设计解决方案，为减轻针对个人和社区的负面言论提供有力的支撑。

文章作者提取的网络结构，包含的数据可以准确表示用户及其关注者之间的联系。网络中的节点将通过历史情绪进行剖析，并通过分类器进行预测，这些分类器是根据研究人员使用动态关键字提取方法收集的数据进行训练的。使用情感网络结构，研究人员可以表征网络中的个人和社区倾向。例如，在#MeToo 运动中的讨论中，节点或社区可能被分类为攻击者、防御者、传播者和旁观者。研究人员还计划提供可能的方法，以鼓励积极讨论，同时减少社交媒体平台上人们所遭受的负面遭遇。

5 启发思考

本文详细介绍了 GloVe 模型，并用它来进行动态关键字的检测。GloVe（全球词表示向量）是一个享誉盛名且常用的词嵌入模型。由于它能够表示数据中的线性子结构，因此备受关注。GloVe 是具有加权最小二乘目标的对数双线性模型，旨在学习单词量，使其点积等于单词共现概率的对数。在产生的单词向量空间中，余弦相似度表示两个单词之间的语义相似度，而向量差异则捕获单词对之间的类比。通过阅读这篇文章，我发现 GloVe 模型在进行动态关键字的检测时是一种非常有力的工具。本文使用 GloVe 模型在各种语料库上训练了 50 维单词向量，其中包括使用 Twitter 的标准搜索 API 获得的 Twitter 数据（允许访问前 7 天内发布的推文）、Wikipedia 数据和 Reddit 数据。在今后的学习和科研中，如果遇到类似的问题，我也可以使用该模型去对抓取的单词进行训练，以精确地得到不同关键字之间的关系，进而判断不同关键字背后不同内容与发言者的相互作用。