应用层读书笔记

B 站讲解（更详细）链接：https://www.bilibili.com/video/BV1XV411k7ru

一、文献信息

1. 作者：Lana Cuthberston, Alex Kearney, Riley Dawson
2. 论文题目：Women, politics and Twitter:Using machine learning to change the discourses
3. 发表途径：AI for Social Good workshop at NeurIPS (2019), Vancouver, Canada.
4. 发表时间：December 2019

二、问题意义

1. 研究问题

为了提高人们对网络暴力的认识，改变女性从政的舆论环境，本文设计、构建并部署了 ParityBOT，一个 Twitter 机器人。它是一个基于人工智能的干预措施，首先建立一个可扩展的模型，通过定量和定性的评估对针对女性从政的恶意推文进行分类和响应，然后通过发送关于有影响力的女性领导人的支持性推文和关于女性在公共生活中的事实来反击针对女性在政治上的辱骂性推文。最后，根据 2019 年阿尔伯塔省和 2019 年加拿大联邦选举期间的干预措施收集的数据，本文分析了 ParityBOT 的影响，其积极作用在公共网络暴力数据集上得到了验证。

2. 研究背景及意义

对于民主政治，参政人员的多样性是必要的。但对于不同人员，参政的机会是不平等的。其中的一个突出是例子就是政治中的性别差异：参政的女性比男性少，这是因为女性在世界各地的政治体系中都面临着更多的障碍，其中之一就是网络骚扰。例如，Twitter 是一个重要的社交媒体平台，可以让政客们分享他们的愿景，并与选民接触。但由于性别歧视，参政的女性在这个平台上受到的骚扰和无理的恶意攻击比男性多得多。

因此，我们应该有所行动，为女性争取平等的权利。本文设计的 ParityBOT 对恶意推文进行分类，并通过发送关于有影响力的女性领导人的支持性推文和关于女性在公共生活中的事实来反击针对女性在政治上的辱骂性推文，最终实现两个目标：1）提高对政治中性别不平等问题的认识，2）积极影响政治中的公共话语。当从政的女性看到 ParityBOT 上对恶意推文的反驳，会感到受到鼓励，并被网络政治社区所接纳；这有助于实现政治中的性别平衡，并改善我们社会中的性别不平等现象。

三、思路方法

1. 分析 ParityBOT 的必要性

在引言部分主要提出了政治体系中女性地位的不平等，并以 Twitter 为例，主要分析了女性参政者在 Twitter 上遭受的网络暴力。由此引出改变这种现状的一种措施，ParityBOT 机器人。

2. 描述 ParityBOT 的技术细节

第 2 部分首先描述了 ParityBOT 如何获取 Twitter 上的推文数据：使用开源 Python 库 Tweepy 作为流侦听器从 Twitter 的实时流 API 收集推文，方法是为每个提到一个或多个感兴趣的候选用户名（女性候选人）的英语推文启动异步分析和存储功能。收集推文过程中，为了避免多次计算同一内容而使分析产生偏差，Tweepy 不跟踪或存储转发的推文。

然后叙述了如何分析推文：使用 Jigsaw 中的透视 API、HateSonar 和 VADER 情感模型作为文本分析模型，将这些模型的输出（即 17 个来自透视 API，3 个来自 HateSonar，4 个来自 VADER）合并到每个推文的单个特征向量中，并将数据储存在表中。在推文分析模型中不涉及任何用户特征，因为虽然这些特征可以提高分类精度，但也可能导致潜在的偏差。经过试验发现，透视 API 的毒性概率指标是对恶意推文进行分类的最好的预测特征，因此在后续的验证中，都选择使用一个透视 API 特性来触发发送积极推文。

之后本文描述了如何建立候选人（参政女性）数据库，以便分析推文的性质；并且叙述了如何收集积极推文进行发送：基本上是来源于志愿者的撰写。

最后，通过对参政相关人士的采访和调查，评估 ParityBOT 带来的社会影响。

3. 分析在两次选举中部署 ParityBOT 的结果

在 2019 年阿尔伯塔省选举和 2019 年加拿大联邦选举部署了 ParityBOT。在两次选举中根据参与者规模的不同设置了不同的阈值。

在阿尔伯塔省选举期间，我们最初将判决阈值设置为毒性评分高于 0.5，以捕获大多数恶意推文，但考虑到我们库中的积极推文数量和 Twitter API 的每日限制，我们发送了太多推文。因此，在 ParityBOT 被激活的前 24 小时之后，我们将决策阈值增加到 0.8，这代表了训练数据中恶意推文的一个显著拐点。考虑到处理推文的数量和速度的增加，我们进一步将加拿大联邦选举的决定门槛提高到 0.9。最终发现，我们发送积极推文的数量是受限的，也是较为合理的。

此外，本文为 ParityBOT 编写了指导方针和价值观，以指导项目的持续发展。例如，ParityBOT 避免了对 Twitter 用户的任何直接“at（@）”提及，防止新一轮网络暴力的出现。

通过在两次选举中对参与者或者其余相关人员的调查，本文发现了 ParityBOT 在警示网络暴力和促进政治公平上的积极影响。

4. 对 ParityBOT 推广的展望

本文希望在更多的国家和地区推广 ParityBOT，以扩大潜在的积极影响。在未来对 ParityBOT 的更新中，系统将会更好地将积极推文与特定类型的暴力推文相匹配，以达到更好的打击网络暴力的效果。

5. 关于 ParityBOT 的详细内容

在附录中介绍了更多关于 ParityBOT 的细节。首先描述了如何将推文进行整理，以便于进行恶意分类；然后简述了文本分类模型划分后推文特征的 24 个具体标签，通过一个实验测试了不同文本分类模型的有效性；之后公布了两次选举时的实验结果；最后附上了研究计划和调查 ParityBOT 影响的具体方法。

四、实验结论

本文设计了一个 Twitter 机器人，ParityBOT，对针对女性从政的恶意推文进行分类和响应，然后通过发送关于有影响力的女性领导人的支持性推文和关于女性在公共生活中的事实来反击针对女性在政治上的辱骂性推文。通过在 2019 年阿尔伯塔省和 2019 年加拿大联邦选举期间部署 ParityBOT，我们发现，对于女性参政者的网络暴力在社交媒体平台上普遍存在，但却难以进行打击，这正在影响我们社区的民主健康和两性平等。ParityBOT 为平等的政治注入希望和积极性，鼓励更多不同的候选人参与。通过使用机器学习技术来解决这些系统问题，我们可以帮助改变舆论环境，将科学进步与人类进步联系起来。

五、启发思考

1. 使用不同模型综合分析

在分析推文时，不仅用到了一种模型，而是将三种模型的分类标签整合在一个向量中，提高了判决的准确性。虽然透视 API 的毒性指标可以被证明是最有效的，但是将三种模型结合可以使得分类更加全面和细致。因此在其他的研究中，不一定要应用某种特定的模型，在计算复杂度可控的范围内，结合多种模型有助于研究准确性的提升。

2. 实验过程中参数的调整

在两次选举的部署时，最初选择的判决阈值不合适，因此进行了修改。我们可以看出，在实验中各项参数并不是一成不变的，而是需要随着实验环境和不同条件的改变进行调整。即使实验条件不变，最初选择的参数也不一定是符合要求的，应该灵活地调整参数以提高实验的准确性。

3. 运用新技术的原则性

使用 ParityBOT 的目标是对抗对女性参政者的网络暴力，对政治公平产生积极影响，而不是对发布恶意推文的用户进行攻击。本文在发布积极推文和反击恶意推文的过程中设置了原则，对社会产生积极影响。一项新技术的出发点往往是积极的，但是若在使用过程中没有规范和原则去约束，最终就可能造成不良影响。

4. 机器学习改变社会

机器学习作为一项技术，不能仅仅限于理论的研究，而是应该应用于实际。本文利用机器学习影响政治公平，对社会产生积极影响，将科学进步与人类进步联系起来。这是科学研究的最终目的和真实价值。任何科研人员都不应该仅限于纸上谈兵，而是应该放眼现实，对人类社会产生积极影响。