信息网络专题研究课文献阅读报告（应用层）

一. 文献信息

1、作者：Lana Cuthberston，Alex Kearney，Riley Dawson

2、题目：Women, politics and Twitter:Using machine learning to change the discourse

3、发表途径：AI for Social Good workshop at NeurIPS (2019), Vancouver, Canada.

4、发表时间：2019

二. 问题意义

1. 研究背景

政治决策中的多样性可以加强民主体制，有利于民主机构的健康发展。然而在加拿大的政治体系内，各级民选政府之间存在性别不平等现象。妇女上任率低，同时面临着系统性的障碍，其中网络平台 Twitter 中发表的针对从事政治活动的女性的仇恨性推文等更加剧了这种不平等现象。

2. 研究问题

本文提出了 ParityBOT，这是一种基于人工智能的，可以直接参与政治社区 Twitter 讨论的工具。它可以通过发表有关女性领导人或者女性公众人物的支持性推文，来反驳那些仇恨性推文，从而更好地影响政治女性的在线言论。

3. 研究意义

Twitter 是政客分享观点和与选民互动的重要社交媒体平台， ParityBOT 通过在 Twitter 上发表有利于提高女性政治地位的支持性推文，减少了政治女性在 Twitter 上公平参与的障碍，使政治女性感到鼓舞。同时还可以提高人们对于政治中性别不平等有关的问题的认识，积极影响政治中的公众话语，有利于在政治中实现性别平衡，并改善社会中的性别平等。

三、思路方法

1. ParityBot 可扩展模型分析

1) Twitter 侦听器

用于定量和定性评估仇恨性推文，可以对针对女性候选人的推文进行收集与分类。使用开源 Python 库 Tweepy ，针对列表每一名候选用户名对实时收集的英语推文进行存储与分析（文本分析模型是在英语语料库上训练的），转发推特不会被跟踪或存储，以免多次分析相同内容来使分析产生误差。实现过程如下：

① 解析推文信息并提取文本

使用正则表达式规则来处理推文（要求符合分析模型）：将文本转换为小写字母，删除URL，删除换行符，用单个空格替换空白，并用文本标签“ MENTION”替换相关页面功能。尽管这些规则可能会使分类器产生偏差，但保证了训练和测试数据集之间的一致性。

② 使用多个文本分析模型对推文评分

首先使用 Jigsaw 的 Perspective API，HateSonar 和 VADER 这三种机器学习情感模型对推文（不包含任何用户特征）可能产生的影响进行评分（0 到 1 之间），三种模型的输出被组合为每个推文的单个特征向量，作为本论文模型的输入，已实现对推文进行分类。在对特征的分析过程中发现，使用单个 TOXICITY 特征几乎与使用所有特征和更复杂的模型一样具有预测性，大规模实施和处理推文也更简单。

③ 将数据存储在数据库表中。

2) Twitter 响应器

对于收集的每条推文，都计算其具有仇恨性的可能性，如果概率高于响应决策阈值，相应地发表积极推文，推文的内容已经过一定程度的审查。

2. 数据集建立

1) 数据收集

① 艾伯塔省 2019 年大选期间：ParityBOT 招募志愿者使用在线资源来创建一个关于所有候选人的数据库，志愿者在该数据库中记录了每个候选人的性别和 Twitter 账号。

② 2019 年加拿大联邦大选期间：ParityBOT 使用 Python 库 gender-guesser 根据每个候选人的姓来预测他们的性别，志愿者尽可能地利用网上收集的确凿证据来手动验证这些预测。

2) 数据验证

首先进行消融实验找到最相关的特征，并使用经过处理的推文数据集 20194 设置了仇恨性质的预测阈值，这些推文被标识为‘hateful’和‘not hateful’。由此数据集由 24 个特征和类别标签组成，将其随机分为训练（80%）和测试（20%）集。接着使用 ADASYN 重新采样并平衡数据集中的类比例，以实现整个数据集平衡（25.4% hateful）。

四、实验结论

1. 实验分析

由特征的正负例上概率分布可得，Perspective API 的 TOXICITY 特征是对推文进行分类的最具一致性的预测特征。

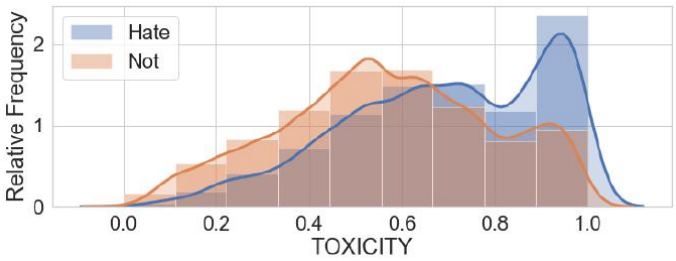

图 1　TOXICITY 特征在正负例上概率分布

通过对训练数据使用了 10 倍交叉验证，进行了消融实验，发现最好的分类器是梯度提升决策树，以及各个特征的相对影响。

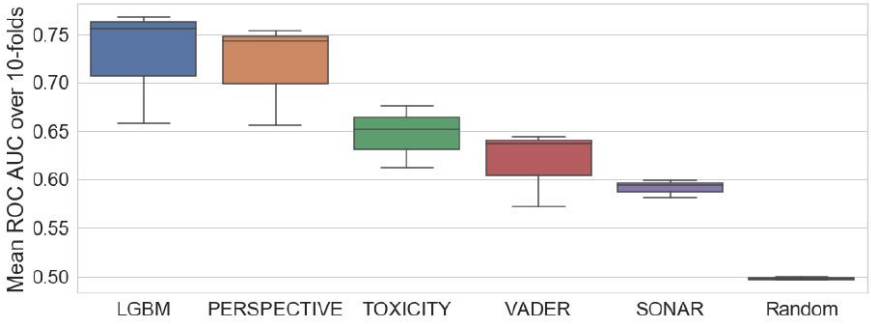

图 2 10 倍交叉验证消融实验接收器工作特性曲线

2. 定量评估

该模型在两次选举过程中发表的积极推特数量均小于推定的仇恨性推特数量。

	Alberta 2019 provincial election Apr. 1-15 2019	**Canada 2019 federal election** Sep. 11-Oct. 26 2019
Total positivitweets sent	973	2428
Total impressions	84,961	304,600
Total retweets	142	529
Total likes	412	1500
Total replies	n/a	30
Total tweets analysed	12,726	228,255
Total tweets scored abusive	1468	9987
Abusive rate	7.65%	4.38%
Total candidates tracked	90	314
Decision threshold	0.8 (80% likely to be abusive)	0.9 (90% likely to be abusive)

表 1 两次选举中推特定量统计

3. 定性评估

通过采访参与政府工作的五个人来评估系统对社会的影响，有关定性评估的完整讨论指南包括研究目的，目标参与者，谈话内容安排。

结论是 ParityBot 确实发挥了一定作用，强调了仇恨言论在现有的社交媒体平台上十分普遍并且难以消除，并且正在影响我们社区的民主健康和性别平等，希望可以积极注入政治，以鼓励更多不同的候选人参加。

五、启发思考

第一次听说可以通过使用机器学习技术来解决一些系统性问题甚至政治问题，以此帮助改变将科学进步与人类进步，感受到了 AI 技术渗透领域之广泛，作用之强大。通过此篇文章的阅读，我拓展了对于人工智能技术的认识，也认识到了知识技术发展对于改善人类社会生活的重要意义。