推特平台2020年美国大选预测报告

2020年11月1日更新

周振坤

首都经济贸易大学统计学院

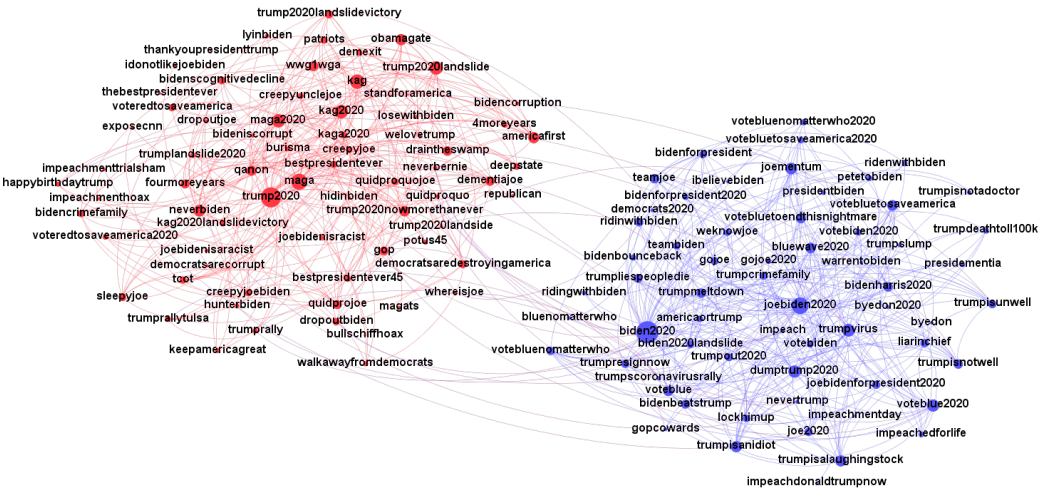
沃民高新科技（北京）股份有限公司

纽约城市大学（City University of New York）

本文使用大规模来自Twitter平台的社交网络数据和机器学习算法以推断用户的政治观点和意见，进而建立选举预测模型，预测2020年美国大选结果。

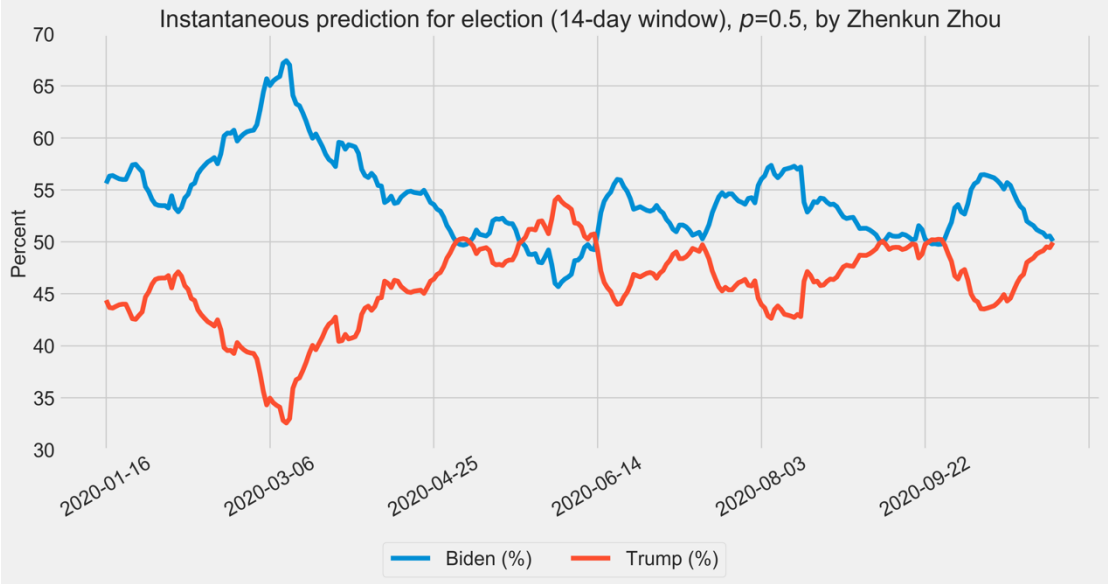
首先确定2020年美国选举的候选人相关的关键词，以收集了大量与美国选举相关的推文，关键词具体包括：trump OR donaldtrump OR realdonaldtrump、biden OR joebiden，这其中主要包含了两个总统候选人名字及其Twitter昵称。截止至2020年11月1日共获取超过七亿条推文数据和二千万用户数据。

利用已经标记的话题标签网络（见下图）对数据集进行标注。若推文中的话题标签只属于一类（特朗普或拜登），那么该推文将自动被标记为该类。对2020年1月1日到2020年8月1日的所有相关推文进行自动标注，最终获取了标记有用户意见的推文训练数据集，推文数据集达到300万。



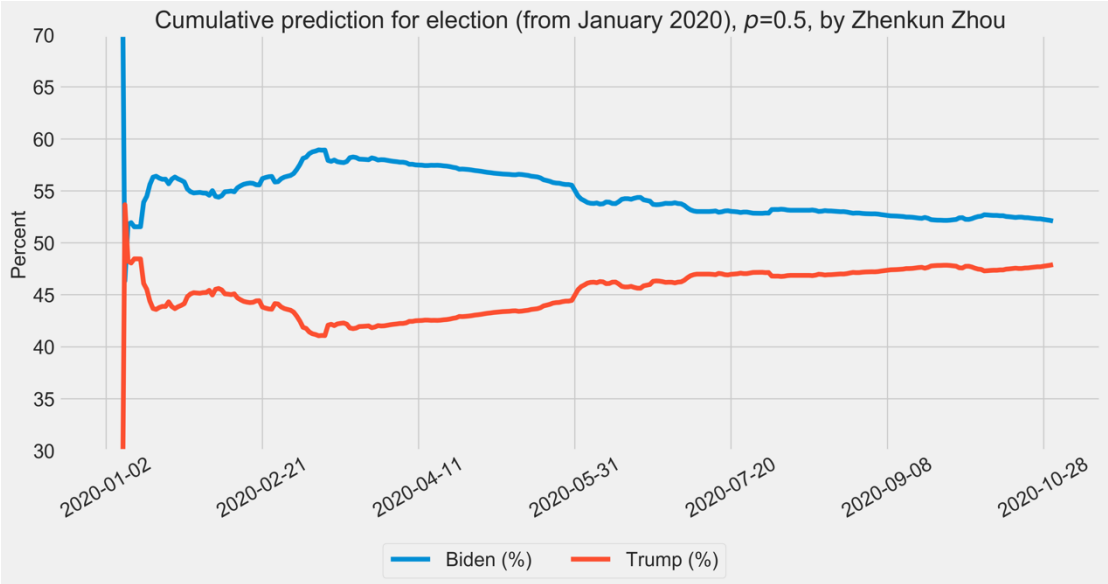
将其中 80%数据作为训练数据，20%作为测试数据。在训练数据中，统计所有文本的一元特征（如单词与标点符号）和二元特征（如二元短语），计算每个文本特征的 TF-IDF 值，选择其中前 100 万个特征作为模型输入特征。通过比较逻辑回归模型（LR）、支持向量机模型（SVM）、随机森林模型（Random Forest）和梯度提升树（GBDT）模型，最终采用逻辑回归模型和 L2 正则化方法为用户意见分类的机器学习模型。经过训练，得到了一个高效的用户意见模型，在两个类别上均表现不俗，平均 F1-score 达到 86%。在训练数据集中进行 10 折交叉验证也达到了 86%的准确率，这表明本文获得了一个很可靠的用户意见分类模型，为选举预测模型提供了坚实基础。

在方法上，本文为公众意见趋势制定了两种不同时间模式的预测，分别为短期窗口预测和历史累计预测。在本实验中，将 w=14 天设置为短期预测事件窗口。大选预测的短期预测结果见图，特朗普截止 11 月的支持率为 49.94%，拜登为 50.06%。



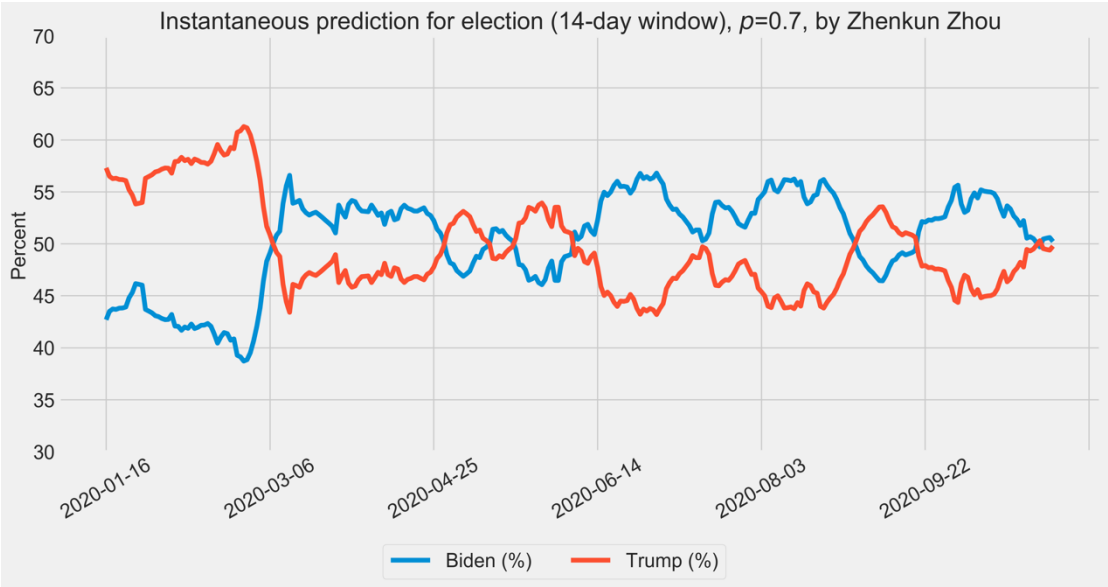
瞬时预测（窗口为 14 天）

本文不仅在一个短期时间窗口中追踪人们对于某候选人的行为和意见，而且追踪和监测社交网络用户在选举前数月的历史累积行为和意见，对于候选人的观点趋势会变得更加明显。历史累计预测结果如下图。特朗普截止 11 月的支持率为 47.90%，拜登支持率为 52.10%。由以上分析结果可得在推特平台上更多民众支持拜登胜利，对拜登支持率更高。本报告预测拜登年将最终赢 2020 年美国大选。



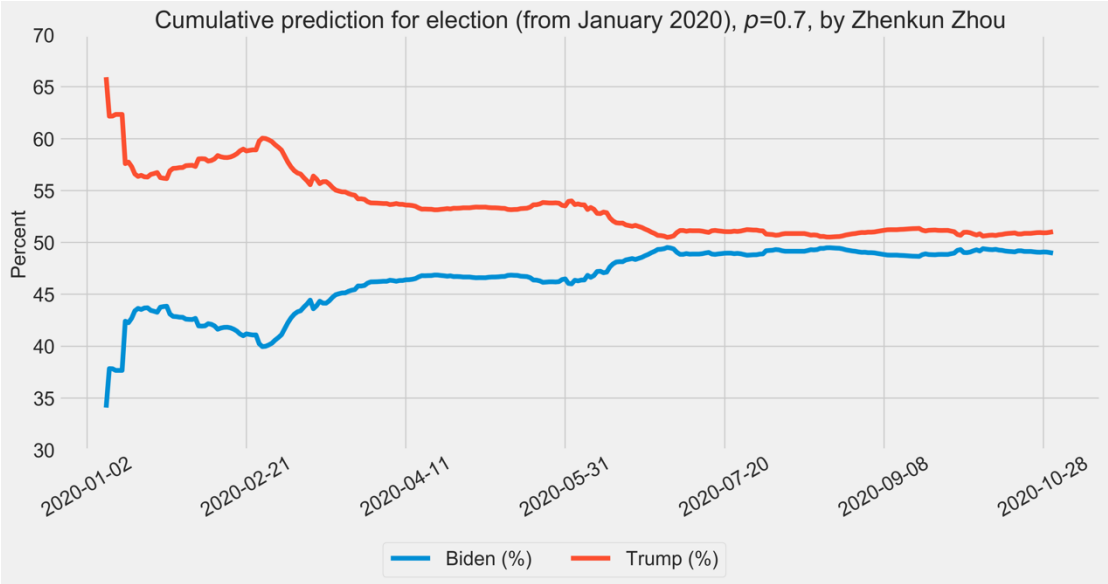
历史累计预测（意见累计自 2020 年 1 月）

除此之外，若提高机器学习模型阈值，即对于用户意见的“敏感”程度，设置模型分类概率阈值为 p=2/3 或 p=7/10，则结果产生反转，表明特朗普支持者的意见（语义上）更加具有极性，支持者选举意向更加坚定。下图为 p=7/10 时，支持率结果。



瞬时预测（窗口为 14 天），p=7/10

结果显示特朗普支持率 49.78%，拜登 50.22%



历史累计预测（意见累计自 2020 年 1 月），p=7/10

结果显示特朗普支持率 51.04%，拜登 48.96%