Abstract:

We study online political debate by combining a theoretical model with a large-scale dataset. Our model explores how users strategically employ aggressive language ("hate speech") and corroborating evidence in an environment of low trust and widespread disinformation. Using a semi-supervised learning approach, we estimate the ideological distance between users in almost 140,000 Twitter interactions. The use of aggressive language in reply tweets is increasing in ideological distance between sender and receiver while the use of evidence (hyperlinks) follows an inverted U-shape. These patterns align with our model's predictions and offer new insights into how online debate works and can (not) be improved.