American Politicians Diverge Systematically, Indian Politicians do so Chaotically:
Text Embeddings as a Window into Party Polarization

ICWSM-2021
June 7-10, 2021

Amar Budhiraja, Ankur Sharma, Rahul Agrawal, Monojit Choudhury, Joyojeet Pal

Microsoft

Introduction

- Demonstrating the use of representation learning as a window into political dialogue on social media through the tweets by politicians on Twitter.

Count/Country	USA	India
Politician Handles	4422	13111
Tweets	2767344	5637474
State Annotations	1500	13111
Party Annotations	598	13111

Table 1. Dataset Statistics

Proposed Method

- Given a politician and a tweet they have written, predict 'K' sequence words from the tweet with the politician ID as input.

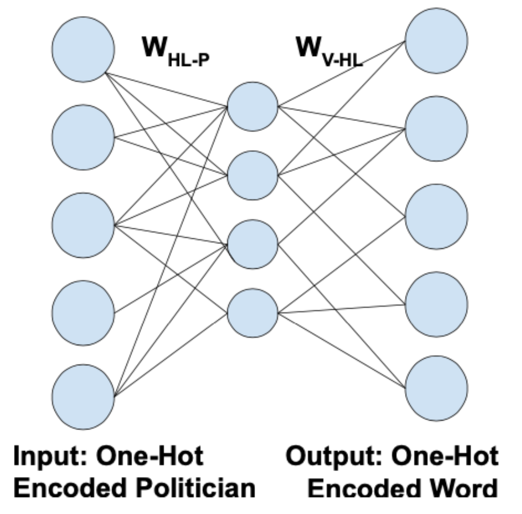

Figure 1. Model Architecture to learn political embeddings

$$\mathcal{L}(\theta) = \frac{1}{M}\frac{1}{K}\sum_{j=1}^{M}\sum_{i=1}^{K}\log(Pr(w_i, P_{t_j}) \mid \theta)$$

Pr -probability of predicting a certain word out of the K sampled
M - total # of tweets
θ - model parameters

Party Prediction Task

- USA: Strong correlation between the word distribution and the predicted party (logistic regression on embedding) of a politician

Method	Accuracy	F1-Score
Proposed Method	**0.9795 (0.0066)**	**0.9793 (0.0066)**
User-DBOW (Ding, Bickel, and Pan 2017)	0.9616 (0.0081)	0.9613 (0.0082)
Word2Vec Baseline (Benton, Arora, and Dredze 2016)	0.9131 (0.0169)	0.9126 (0.0170)
TF-IDF Baseline	0.9036 (0.0278)	0.9026 (0.0286)

Table 2. Results of Party Prediction Task for USA Dataset.

Method	Accuracy	Precision	Recall	F1-Score
Proposed Method	**0.8518** (0.0100)	**0.9206** (0.0061)	**0.6062** (0.0160)	**0.6990** (0.0146)
User-DBOW (Ding, Bickel, and Pan 2017)	0.8151 (0.0156)	0.9053 (0.0083)	0.5712 (0.0168)	0.6651 (0.0163)
Word2Vec Baseline (Benton, Arora, and Dredze 2016)	0.7677 (0.0143)	0.8533 (0.0067)	0.5004 (0.0159)	0.5756 (0.0160)
TF-IDF Baseline	0.6527 (0.0238)	0.8110 (0.0127)	0.4745 (0.0274)	0.5315 (0.0233)

Table 3. Results of Politician Polarization for Indian Dataset.

- India: Party prediction results are not as accurate as of the USA. Possible factors beyond the party of a politician influence their writing on social media

Political Affiliation and Topical Communities

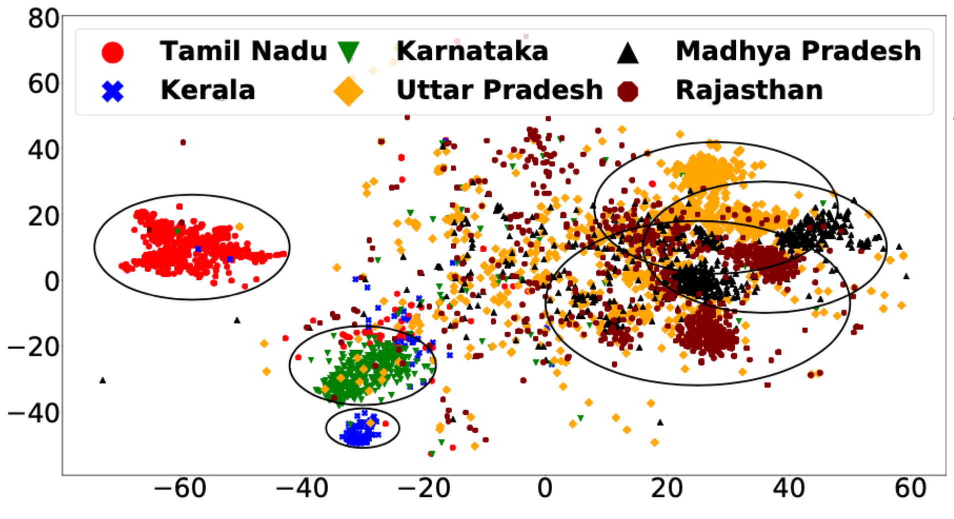

Figure 2. Scatter plot of embeddings for six states in India

- India: Distinct bifurcation between the Northern (BJP-majority) states and the Southern (Regional & Ethnolinguistic) states; Geography (state) is also a crucial factor

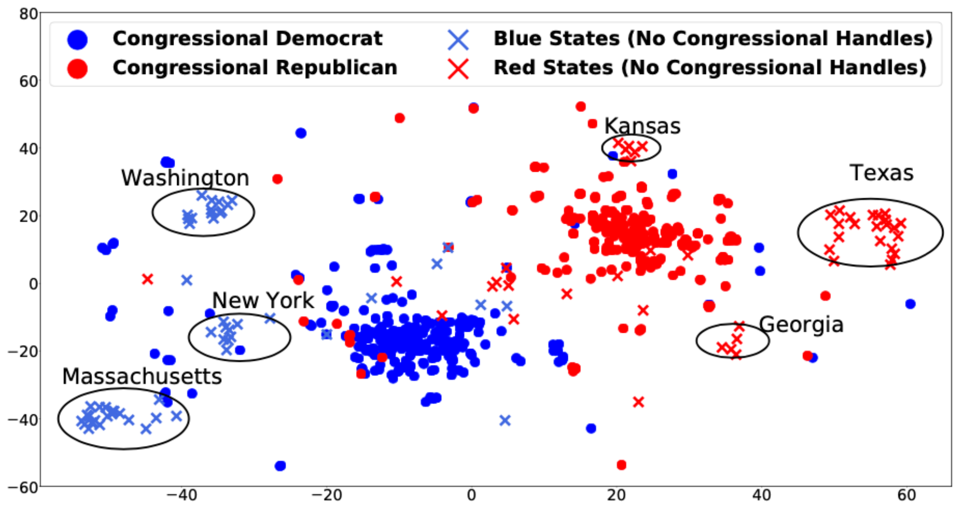

Figure 3. Scatter plot of the embeddings for USA

- USA: Clear party-based affinity from state politicians as well as congressional handles. Also, clear democratic and republican clusters within each state leading to high party prediction accuracy

Discussion

- USA: Republicans and Democrats sit virtually on two separate planes, closely mirror that separation in politicians' online discourse
- India: Hindi-speaking northern states are highly overlapping, representing the hegemonic control of the ruling BJP, while the non-Hindi speaking states are strikingly separated into their own clusters.
- We demonstrate that representation learning gives us a window into how polarized our politics are, and think about the risks that bring.