

Estimating Ideology for U.S. House of Representatives Primary Candidates: Word Embeddings and Campaign Websites

Colin R. Case

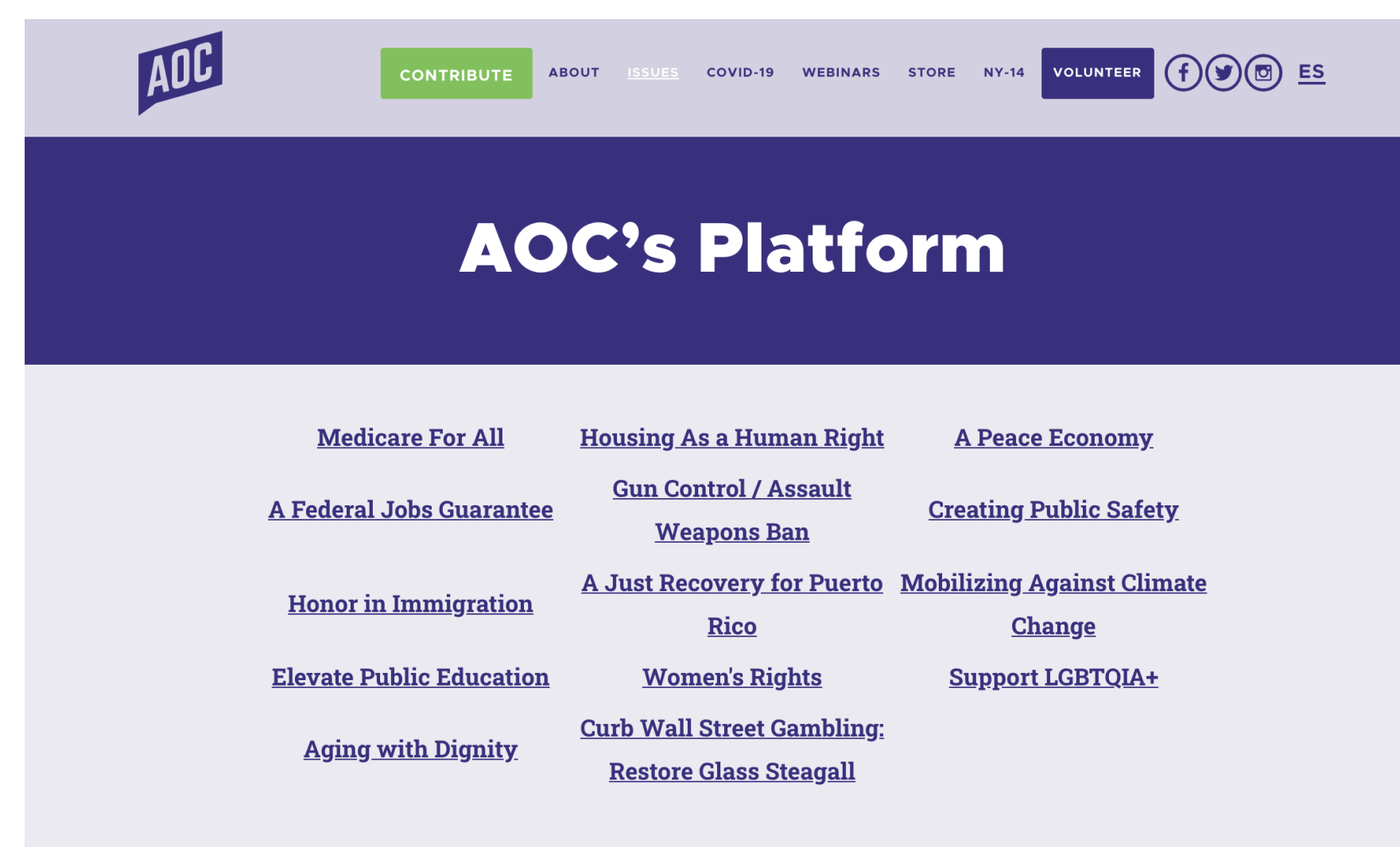
University of North Carolina; crcase@live.unc.edu

Research Objective

Can candidate issue platforms produce conceptually valid ideology estimates?

Data: Candidate Websites

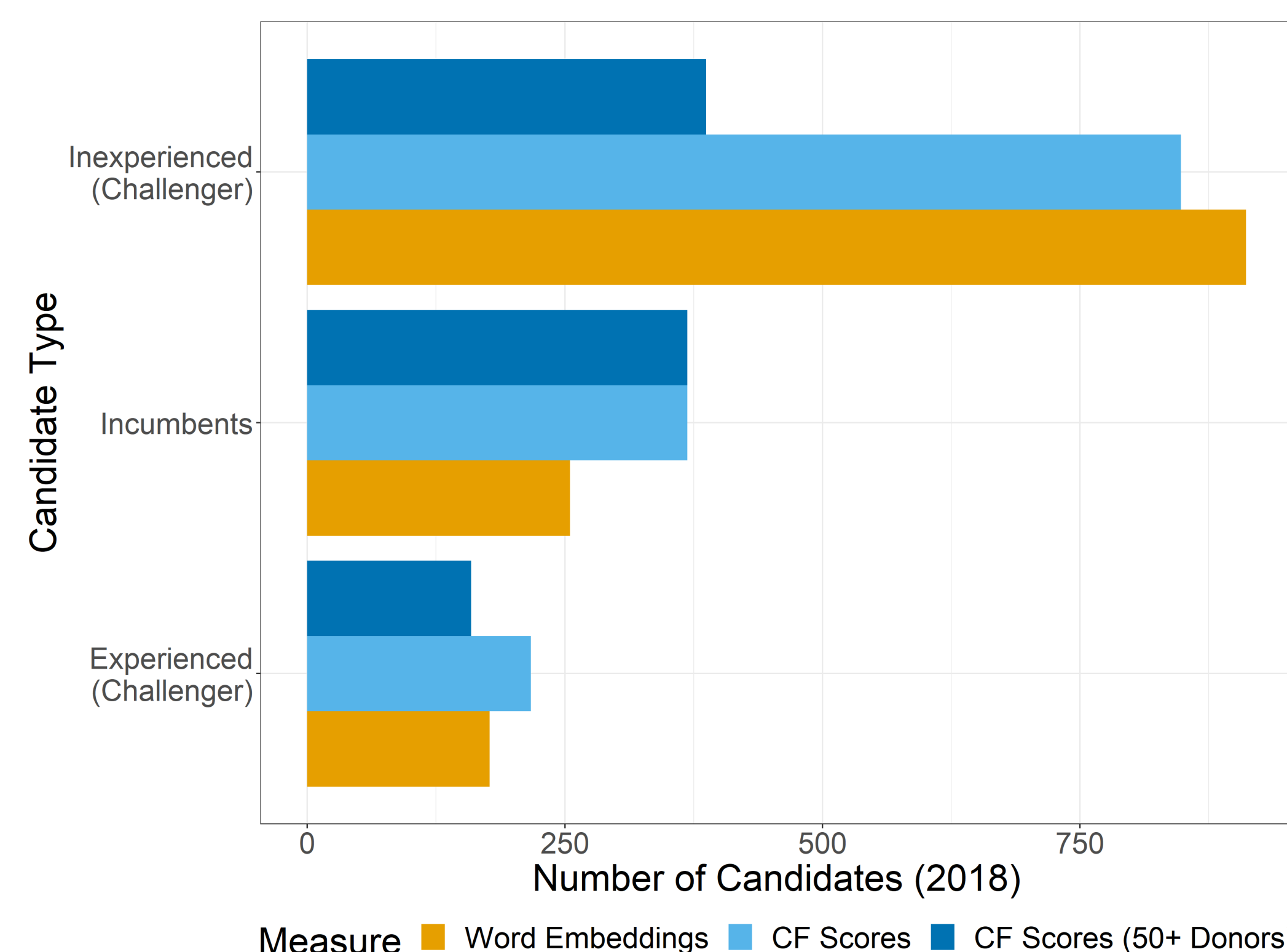
2018 and 2020 House of Representative primary candidates' (N = 2,707) issue platforms (Porter, Treul, and McDonald 2021).



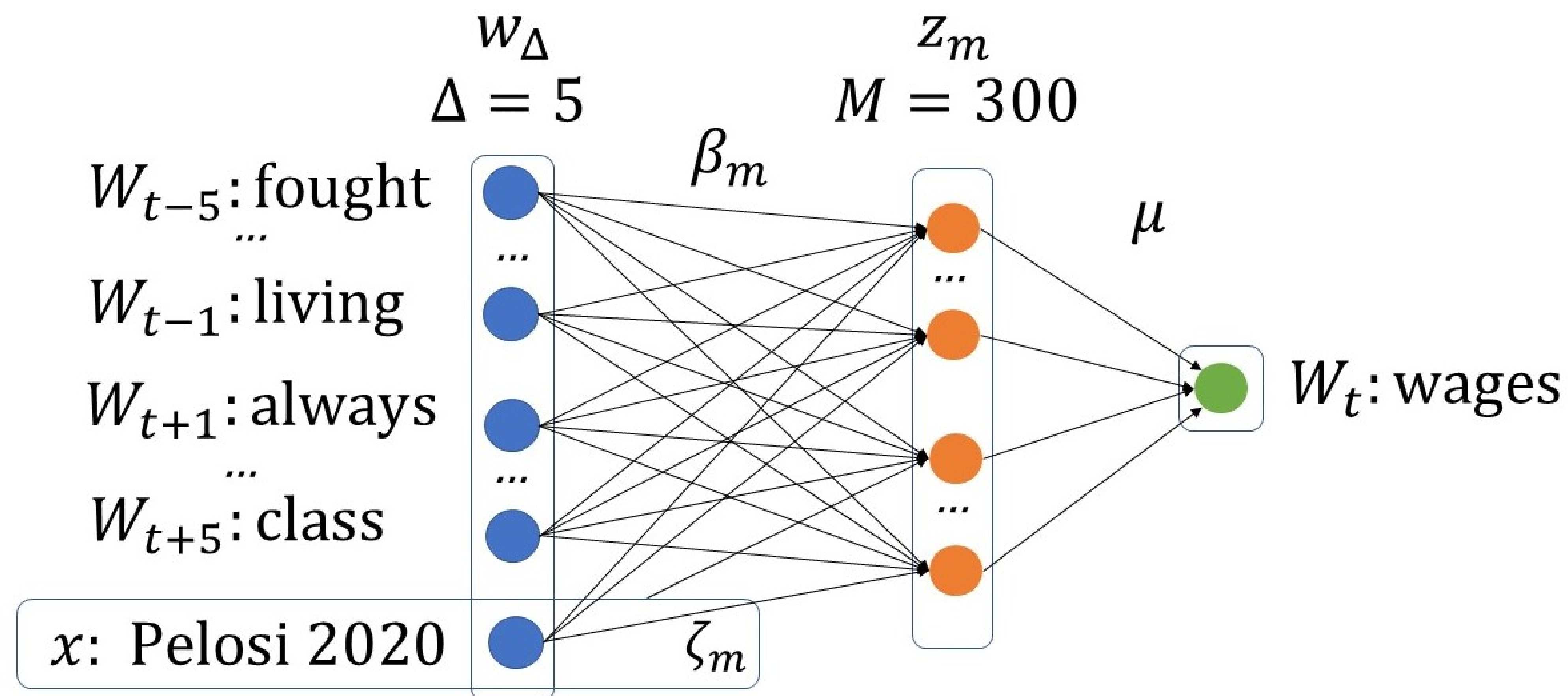
Data Advantages

- Increased coverage of inexperienced candidates versus prior measures
- Captures candidates' behavior instead of donors' perception of candidates
- Lower influence of partisanship on underlying data

Number of Candidates Included: 2018 Primary Elections

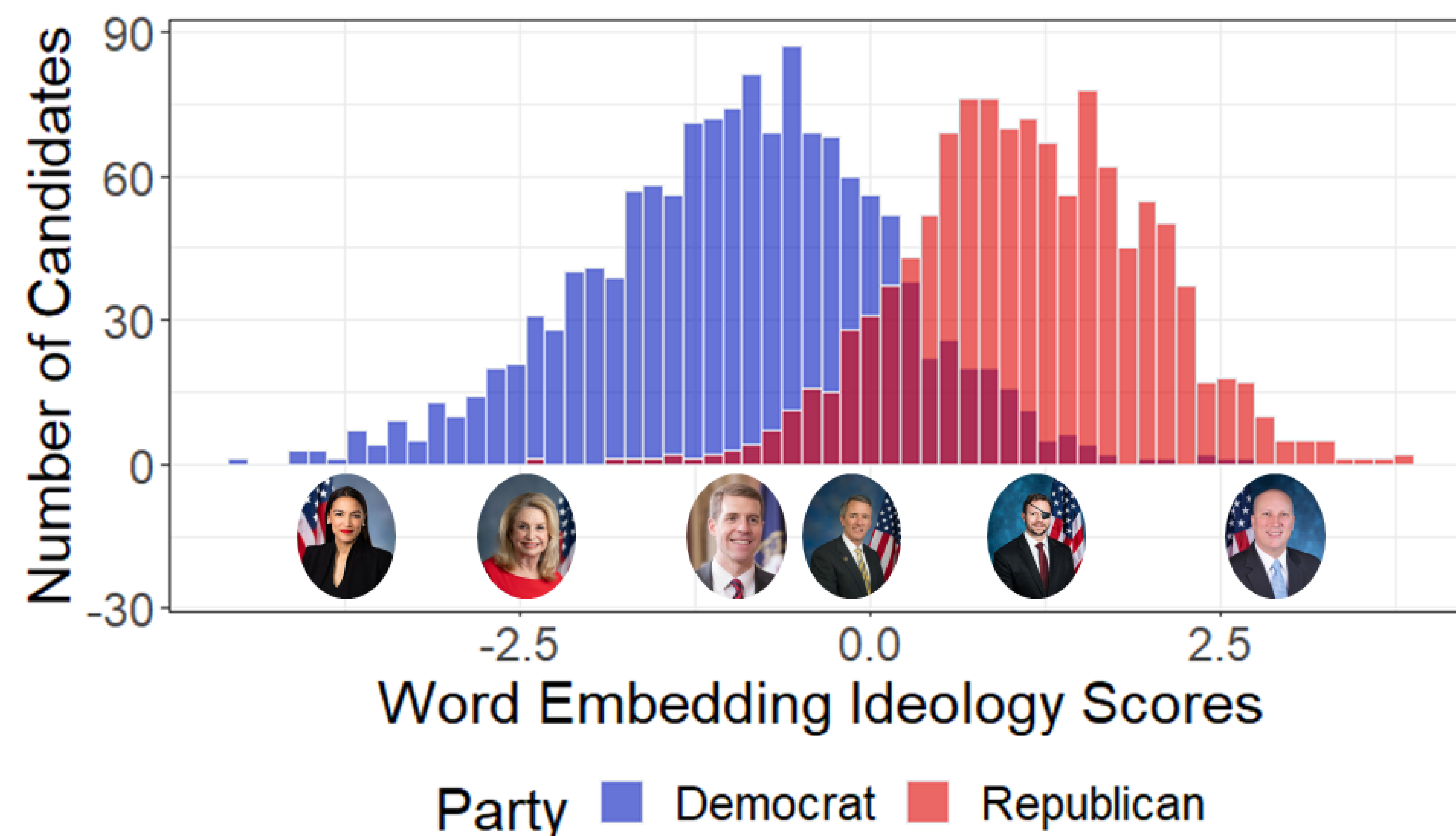


Method: Word Embeddings



- Input variables, w_Δ , where $\Delta = 5$
- Hidden node, z_m , where $M = 300$, $z_m = f(w'_\Delta \beta_m + x' \zeta_m)$ and f is the average
- Document level vector, x , for each candidate-year combination.
- PCA to reduce dimensions of resulting vector $\zeta_{m,i}$ for each candidate-year, i .
- Ideology scores are calculated by adding first two dimensions.

Word Embedding Ideology Scores



Polarization Comparison



Re-scaled to have mean of 0 and standard deviation of 1. Word Embeddings include all 2018 and 2020 candidates, CF Scores include 2018 candidates, and Nominate includes all members from the 116th and 117th. Recipient CF Scores from Bonica (2018). DW-Nominate scores from Lewis et al. (2022).

Correlations with "Gold Standard" Ideology Estimates

Measure	All	Dem	Rep
CF Scores	0.70	0.03	0.21
CF Scores (50+ Donors)	0.80	0.15	0.41
DW Nominate	0.79	0.15	0.27

Conclusions

- Word Embeddings produce conceptually valid ideology estimates of candidates
- Potential to better capture changes in campaign behavior across election cycles
- Intra-party polarization is under estimated and the distance between parties is over estimated with prior measures