Reinforcement Learning: Do Politicians Optimize for Engagement on Twitter?

Abhi Ramaswamy | PhD Candidate, Stanford University

Motivation

- Literature: social media encourages politicians to be polarizing because such content receives greater engagement (e.g., Ballard et al. 2022; Brady et al. 2017; Finkel et al. 2020)
- Assumes that campaigns adjust their content based on what gets attention.
- We test this assumption and whether this behavior varies by ideology, district competitiveness, and incumbent status.

Data

- Tweets by **all** congressional candidates (House and Senate, challengers and incumbents) in the 2020 and 2022 election cycles.
- Exclude retweets, challengers receiving less than two percent in their primary election, and candidates sending less than 50 tweets in a given election cycle.
- Totals: 3.7 million tweets sent by 1,832 candidates (1,209 challengers and 623 incumbents)

Setup

RQ: does going viral on tweet k increase the likelihood of tweeting similar content on subsequent tweet $k+1$?

Estimand:

$$ATT = \sum_{i=1}^{N} \tau_i w_i,$$

where for the general kth tweet sent by candidate i (chronologically ordered),

$$\tau_i = \mathbb{E}[\text{similarity}(k+1,k) \mid k \text{ goes viral}] \\ -\mathbb{E}[\text{similarity}(k+1,k) \mid k \text{ not viral}],$$

the within-candidate effect of going viral on the similarity of future content to a tweet. Where V_i is the number of viral tweets sent by candidate i,

$$w_i = \frac{V_i}{\sum_{j=1}^{N} V_j}.$$

Defining Virality: Tweets receiving $\geq 100\%$ increase in engagement relative to a candidate's month-level mean, where engagement = retweets + likes.

Sentence-Embedding Matching Estimator

Approach: Within-candidate, match each viral tweet k to the non-viral tweet $\tilde{k}$ with highest sentence-embedding similarity (computed with SBERT) to k.

Specification: For candidate i, tweet k, and window b, estimate the following 2FE regression

$$\bar{Y}_{i,k,b} = \hat{\tau} D_{i,k} + \gamma_i + \eta_t + \epsilon_{i,k,b} \quad (1)$$

where

$$D_{i,k} = \mathbf{1}\{k \text{ is viral}\},$$

and

$$\bar{Y}_{i,k,b} = \frac{1}{b}\sum_{j=1}^{b} \text{embed\_similarity}(k+j,k)$$

is the average embedding similarity to tweet k among the b tweets following k.

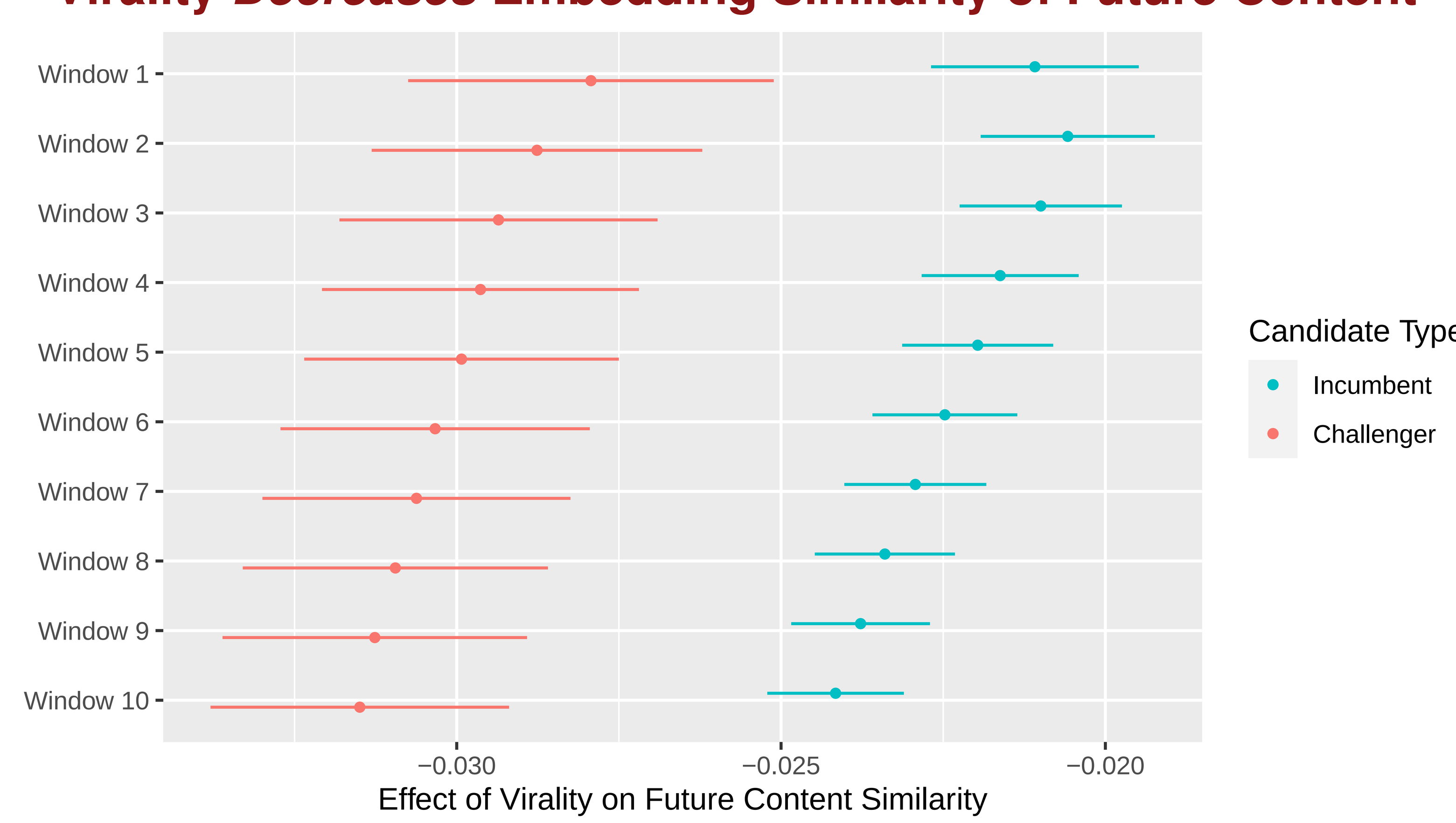
- Candidate fixed effects γ_i provide within-candidate comparisons. Week fixed effects η_t account for temporal patterns correlated with both virality and content similarity.
- Estimate separate models for challengers and incumbents, and for each $b \in \{1,2,\dots,10\}$ to allow for flexible update times.

Identification Assumption: Among matched pairs $(k,\tilde{k})$ with sufficiently high embedding similarity, virality is as-if random:

$$\bar{Y}_{i,k,b}(0) \perp D_{i,k} \mid \text{embed\_similarity}(k,\tilde{k}) > \delta$$

We assume that $\tilde{k} = \text{argmax}_{k' \neq k} \text{embed\_similarity}(k',k)$ satisfies this condition.

Virality *Decreases* Embedding Similarity of Future Content



- The effect grows with window size — after a viral tweet, candidates progressively diverge from prior content relative to how they would have behaved absent virality.

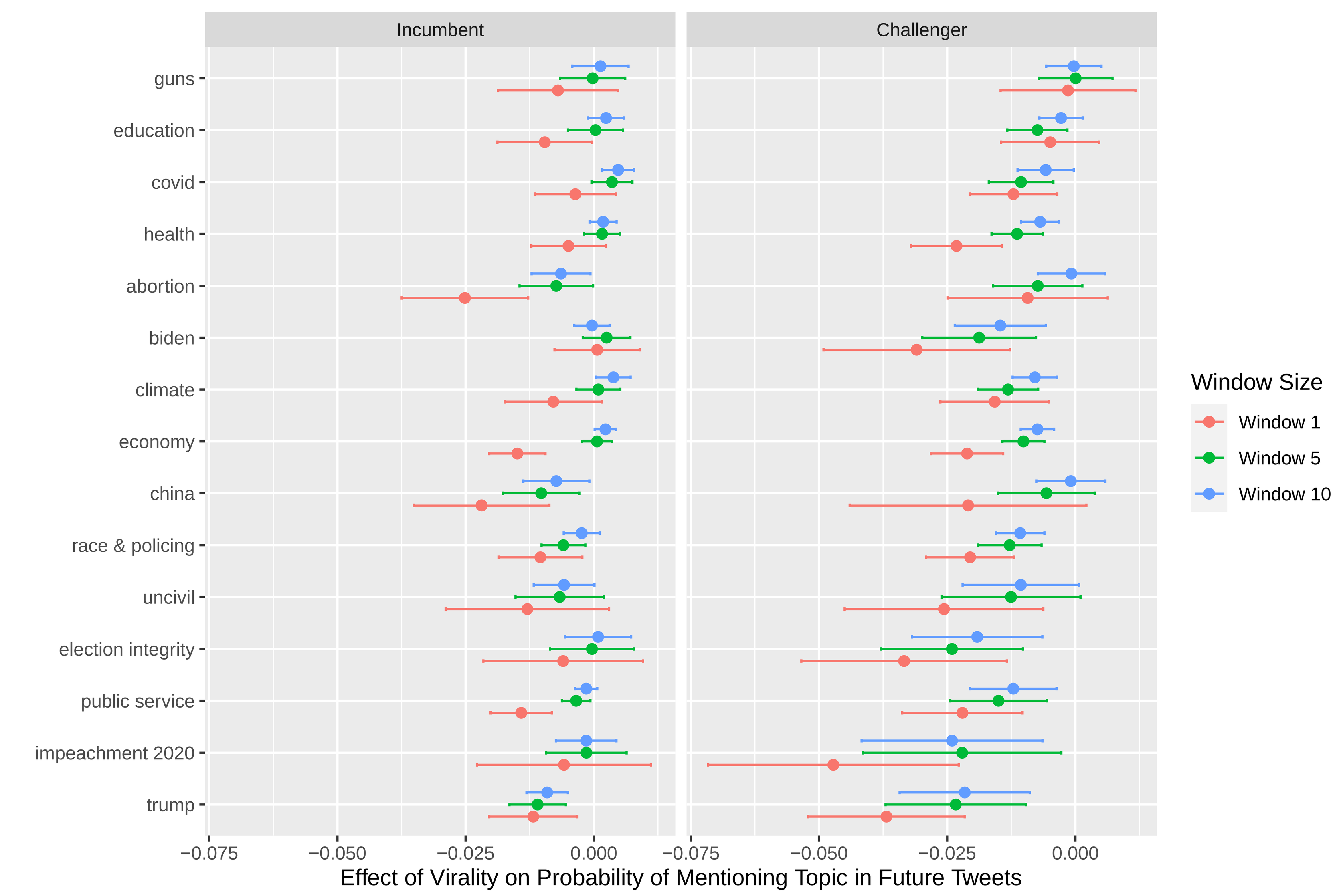
Estimating Topic Optimization

- Classify tweets into 15 topics using dictionary method (validation stats)
- **Specification**: For each topic m, estimate (1) on the subset of tweets that mention m. For each such tweet k, define the outcome as

$$\bar{Y}^{(m)}_{i,k,b} = \frac{1}{b}\sum_{j=1}^{b} \mathbf{1}\{k+j \text{ mention } m\},$$

the share of b tweets following k that also mention m.

Candidates Less Likely to Discuss a Topic after Going Viral with it



Optimization Unrelated to Extremity or District Safety