Social Media Inference & Manipulation

Dr. Yusuf Mucahit Cetinkaya

Postdoctoral Scholar
The University of Edinburgh

What is social media manipulation?

- Manipulation[1]:
  - gen. The action or an act of manipulating something; handling; dexterity. Also (occasionally): the making of hand motions. (1801–)
  - The action or an act of managing or directing a *person*, etc., esp. in a skilful manner; the exercise of subtle, underhand, or devious influence or control over a *person, organization*, etc.; interference, tampering. (1828–)

[1] Oxford University Press. (n.d.). Manipulation, n. In Oxford English dictionary. Retrieved May 1, 2025, from https://doi.org/10.1093/OED/1109487533

What is social media manipulation?

- Manipulation[1]:
  - gen. The action or an act of manipulating something; handling; dexterity. Also (occasionally): the making of hand motions. (1801–)
  - The action or an act of **managing** or **directing** a *person*, etc., esp. in a **skilful manner**; the exercise of subtle, underhand, or devious **influence** or **control** over a *person*, *organization*, etc.; **interference**, **tampering**. (1828–)

[1] Oxford University Press. (n.d.). Manipulation, n. In Oxford English dictionary. Retrieved May 1, 2025, from https://doi.org/10.1093/OED/1109487533

What is social media manipulation?

- Manipulation[1]:
  - gen. The action or an act of manipulating something; handling; dexterity. Also (occasionally): the making of hand motions. (1801–)
  - The action or an act of **managing** or **directing** a *person*, etc., esp. in a **skilful manner**; the exercise of subtle, underhand, or devious **influence** or **control** over a *person*, *organization*, etc.; **interference**, **tampering**. (1828–)
- Social media manipulation:
  - The action or practice of **skilfully managing** or **influencing** *users*, *discourse*, or *perceptions* on social media platforms, esp. through subtle, underhand, or devious means; the strategic **interference** with online content, visibility, or engagement to handle opinions, behaviours, or outcomes.

[1] Oxford University Press. (n.d.). Manipulation, n. In Oxford English dictionary. Retrieved May 1, 2025, from https://doi.org/10.1093/OED/1109487533

What is social media manipulation used for?

How is social media manipulation carried out?

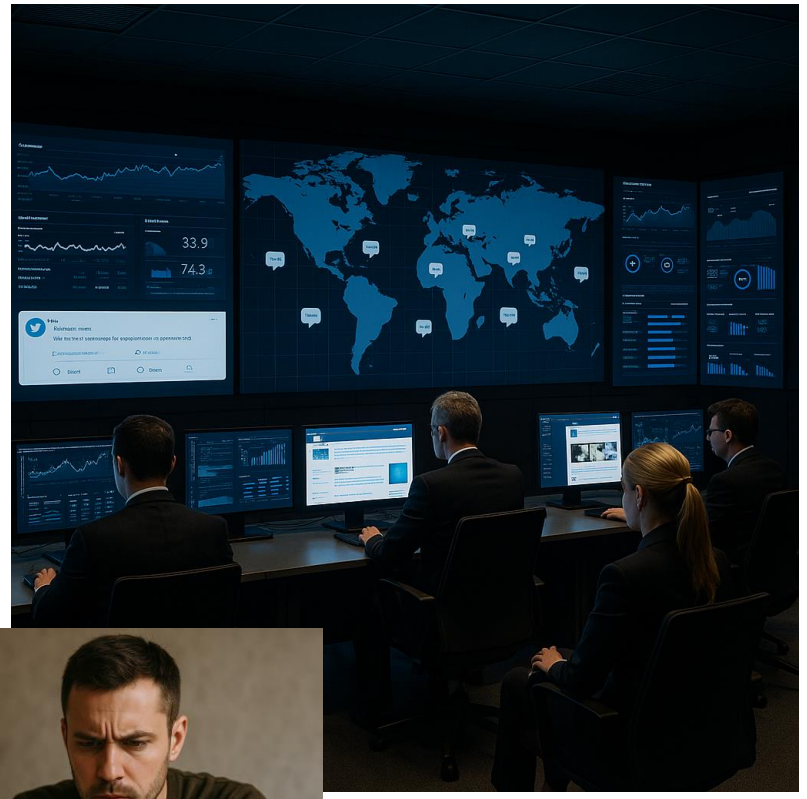
- **Disinformation:** Spreading false or misleading content.
- **Astroturfing:** Simulating grassroots movements through fake accounts.
- **Troll farms & bots:** Coordinated inauthentic activity for disruption or amplification.
- **Deepfakes & manipulated media:** Altered visuals to deceive audiences.
- **Algorithmic exploitation:** Gaming ranking systems to promote content.

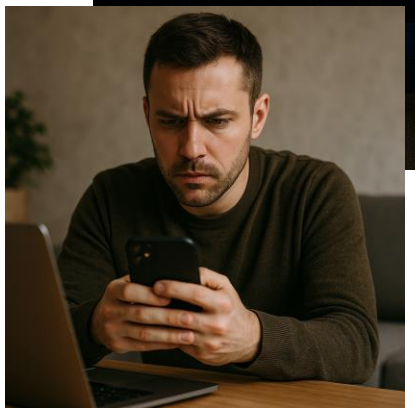
… and in some other ways to be studied

Why is social media manipulation important?

Why should we care?

- **Existence of Bad Actors:** Malicious entities exploit social platforms with hidden agenda.
- **Manipulation Detection:** Detecting a manipulation is not a precaution.
- **Censorship Limits:** Bans galvanize the conspiracy theories.
  It is not scalable for bureaucracy & justice.
- **Platform moderation:** The conveyed message finds another way to be distributed.
- **Black-Hat vs. White-Hat:** Ethical vs. unethical manipulation mirrors hacker culture.
- **Benevolent Manipulation:** Fight back with disinformation and polarization same way.



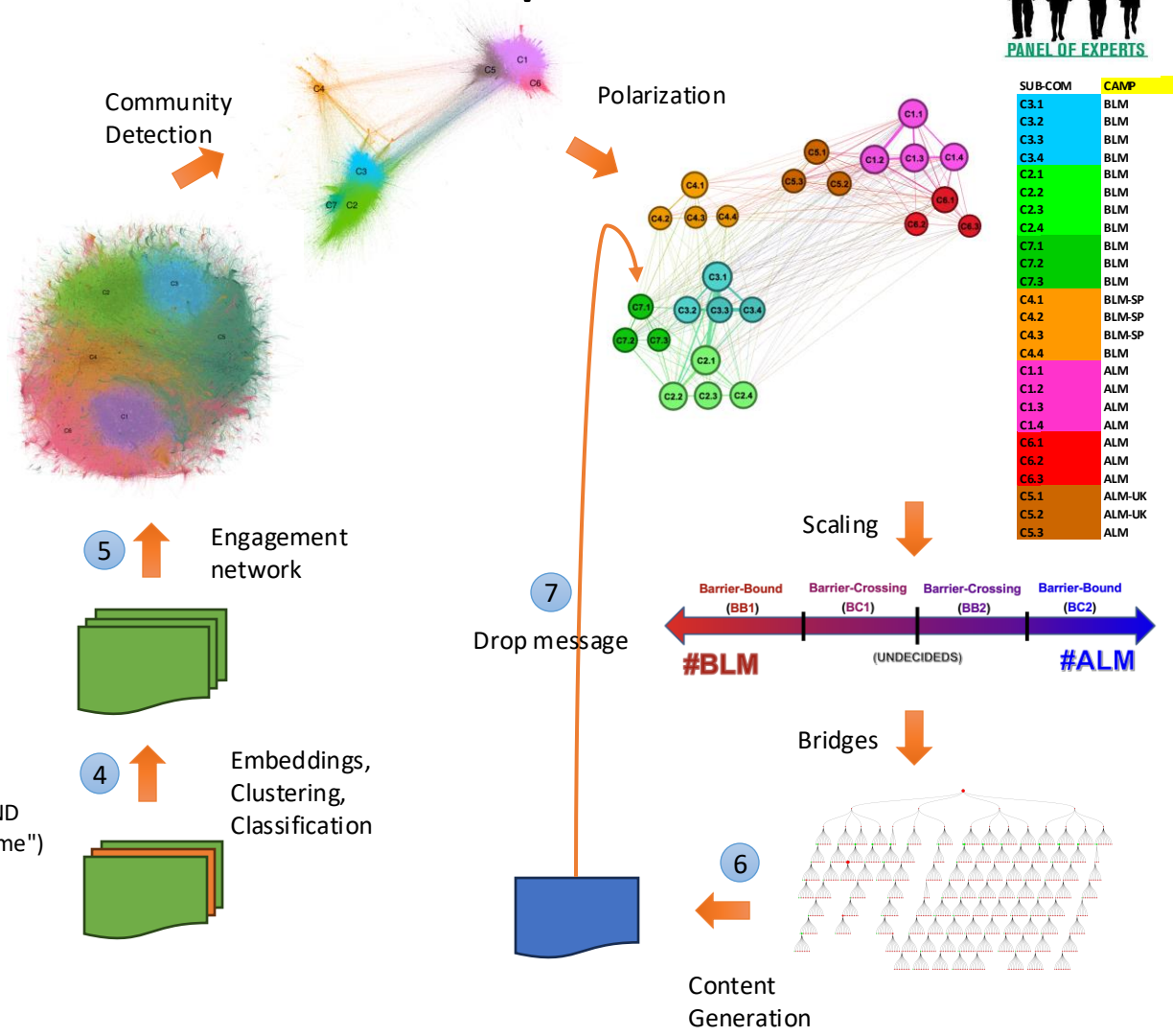


Is benevolent social media manipulation ok?

tl;dr: controversial and risky; in the grey area.

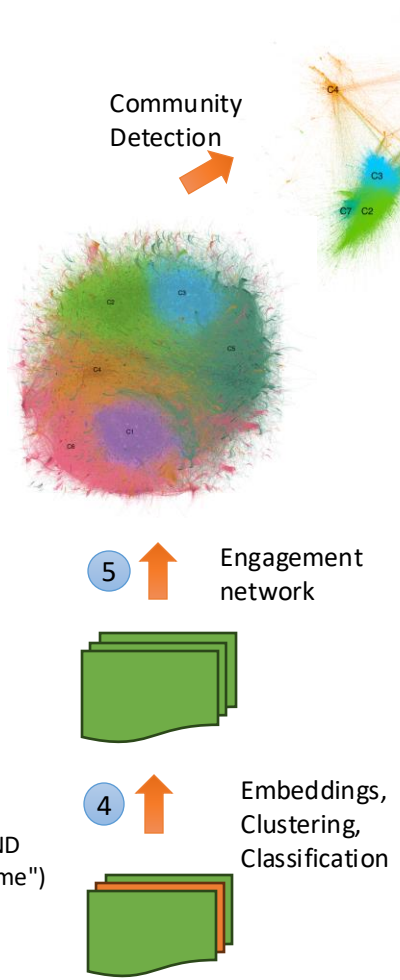
Required steps for social media manipulation

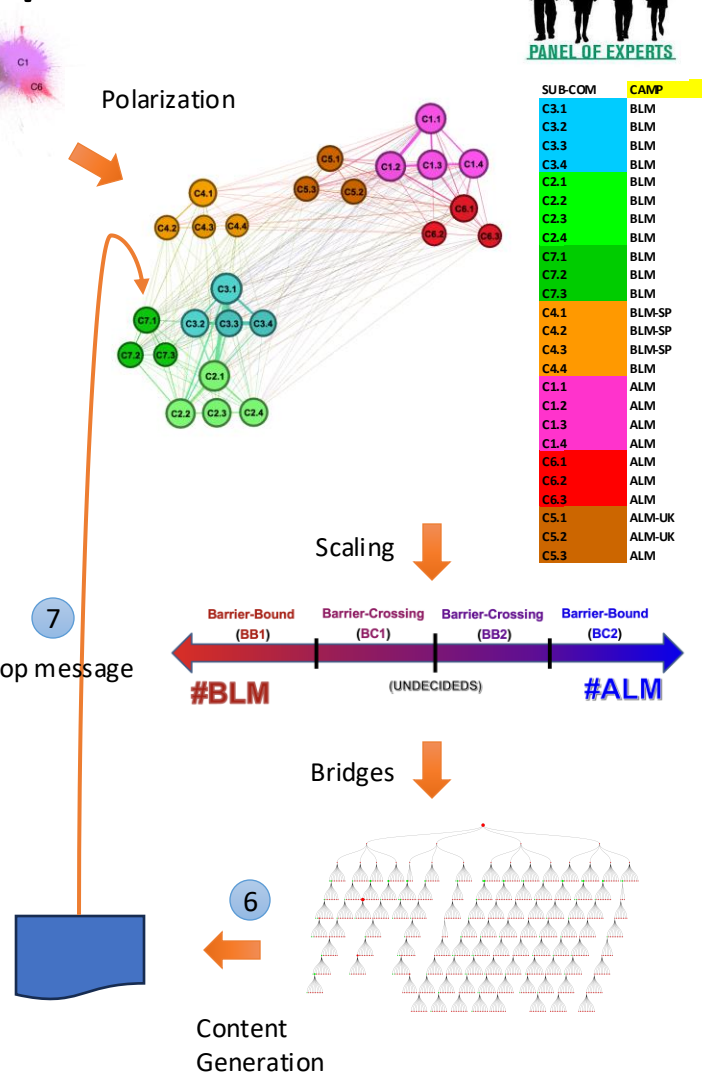
1. Get your target sample
2. Structure their vocabulary
3. Monitor social media via hunting queries
4. Clean the collection to keep related content
5. Understand their
   a) Communities
   b) Language & jargon
   c) Topics
   d) Narratives
   e) Stances
   f) Engagements
   g) Polarization
   h) Scaling
   i) Bridges
6. Generate programmable content
7. Drop your message into the bubble



Required steps for social media manipulation

1. Get your target sample
2. Structure their vocabulary
3. Monitor social media via hunting queries
4. Clean the collection to keep related content
5. **Understand** their
   a) Communities
   b) Language & jargon
   c) Topics
   d) Narratives
   e) Stances
   f) Engagements
   g) Polarization
   h) Scaling
   i) Bridges
6. Generate programmable content
7. Drop your message into the bubble





Leadership Types[2]

- **Barrier-Bound Leaders**
  - Identified as those who seek to advance their group's interests by working **solely** or primarily **within the group.** Outside groups are perceived as irrelevant, polite acquaintances or even as opponents decidedly not to be involved with the pursuit of the group's aims.
- **Barrier-Crossing Leaders**
  - Identified as those who **understand** through observation and conversation what leaders of **other** groups seek to achieve and reciprocally are clear with other leaders about their own group's interests and priorities.
  - They work primarily for their group's interest by **engaging** with members of **other** groups to pool power and resources in identifying common interests, the scope and sequence of tasks to pursue **common interests**, executing **cooperative tasks**, and jointly evaluating their cooperative **effectiveness**.

[2] Buhrmester, M. D., Cowan, M. A., & Whitehouse, H. (2022). What Motivates Barrier-Crossing Leadership?. New England Journal of Public Policy, 34(2), 7.

Cross-Partisan Interactions on Twitter

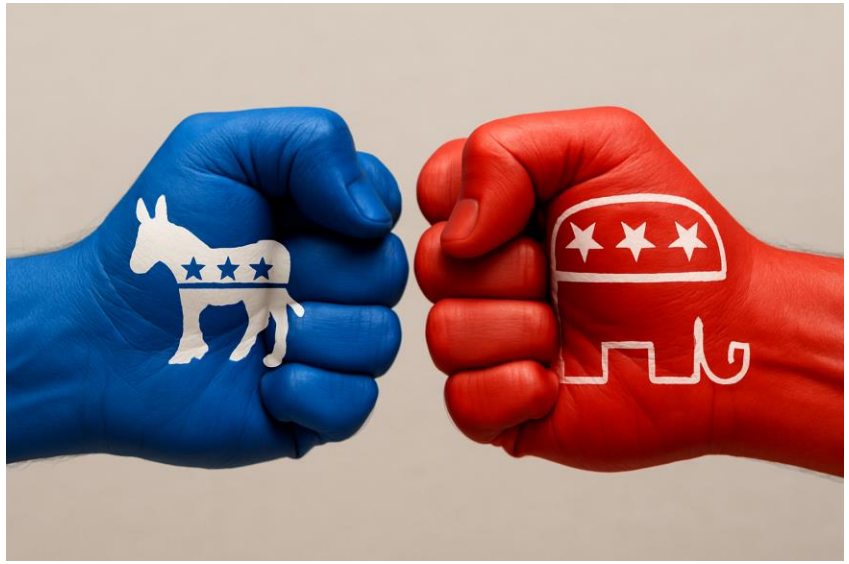
Cross-Partisan Interactions on Twitter

Yusuf Mücahit Çetinkaya[1,2]*, Vahid Ghafouri[3,4,5]*, Guillermo Suarez-Tangil[4], Jose Such[6,7], Tuğrulcan Elmas[1]

[1]University of Edinburgh
[2]Middle East Technical University
[3]Oxford Internet Institute
[4]IMDEA Networks Institute
[5]Universidad Carlos III de Madrid
[6]King's College London
[7]Universitat Politecnica de Valencia

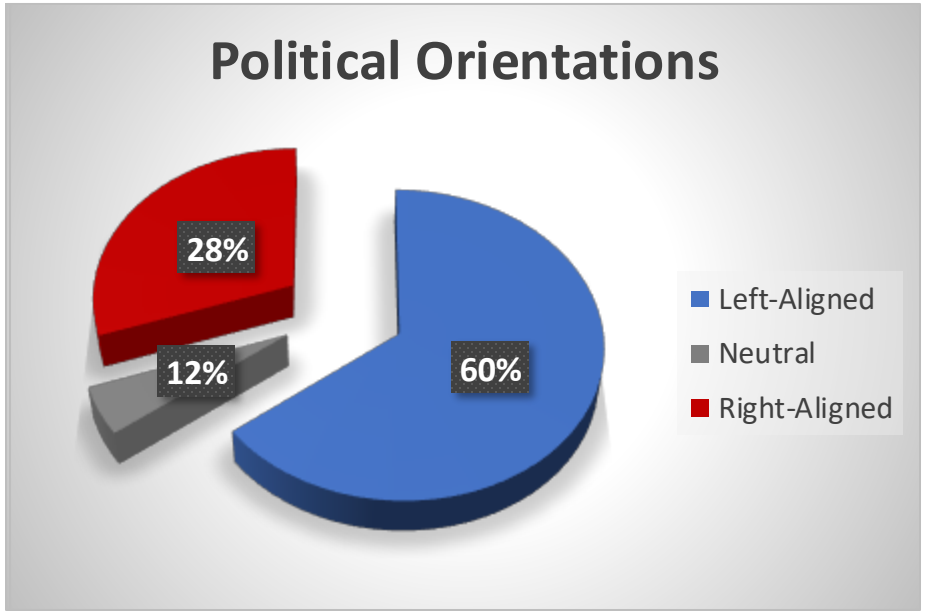
Çetinkaya, Y. M., Ghafouri, V., Suarez-Tangil, G., Such, J., & Elmas, T. (2025). Cross-Partisan Interactions on Twitter. ICWSM '2025.

- When and why do CPIs occur?
- Are they healthy (civil) or toxic (hostile)?
- What do they talk about?
- Can they bridge political divides?
- Is there a tone difference between interactions?



Dataset & Methodology

- 3M+ tweets from 2020 (US context, replies only)
  - 1.8M **direct replies** to 1.2M tweets used
  - **661K CPIs** (~34%)
  - 683K repliers to 211K root authors
- Political orientation from Barberá's method
  - Bayesian inference on users following data
  - **764K authors** have **political orientation** score
  - **494K (60%) left-aligned** vs. **232K (28%) right-aligned**
  - Removed 38K (12%) authors neutral (score: -0.1 to 0.1)
- Toxicity detection: **Perspective API**
  - Gives a toxicity score ranging 0-1
  - Threshold of 0.61[3] is used for defining toxic content
- Stance & sentiment: **Mistral-7B-Instruct** LLM annotations
  - 400K sample for D→D, D→R, R→D, R→R
  - Asked sentiment of the root tweet, stance of the reply tweet as three adj.
  - 97%, 88%, 88%, 82% consistency with human annotation



[3] Kumar, Deepak, et al. "Designing toxic content classification for a diversity of perspectives." Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021). 2021.

RQ1: What are the characteristics of users in CPI?

- **Profile & Tweet Attributes vs. CPI Rates** using logistic regression
  - β_0 is the intercept of the model,
  - β_i are the coefficients associated with the n predictor variables X_i,
  - X_i represent the various author-related metrics included in the model.
- 1.9M+ tweets with their attributes where label=1 if CPI
- 3 models trained (1) all, (2) democrats, (3) republicans

replier_followers	replier_following	replier_tweet_coun t	...	root_like_count	is_CPI

$$\text{logit}(P(Y=1)) = \log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \sum_{i=1}^{n} \beta_i X_i$$

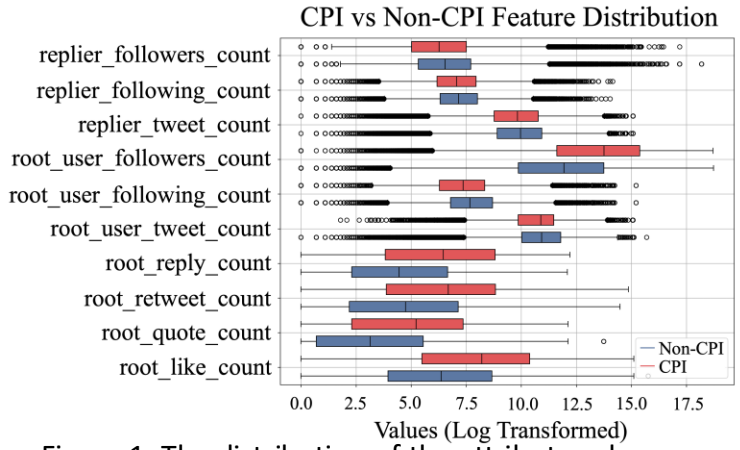


Figure 1: The distribution of the attribute values among the CPIs and non-CPIs.

RQ1: What are the characteristics of users in CPI?

- **Repliers in CPIs** tend to be smaller or newer accounts with fewer followers, followings, and tweets.
- **Root tweets from popular users** are more likely to attract CPIs, especially from Democrats.
- **Reply count** on a tweet is the strongest predictor of CPI, showing a rich-get-richer effect.
- **Likes** increase CPI likelihood for Republicans but reduce it for Democrats.
- **CPIs are more influenced by root tweet attributes** than by replier characteristics.

Variable	All	Dem.	Rep.
const	-0.656	-0.482	-0.672
$\texttt{replier}_{followers\_count}$	-0.025	-0.158	0.020
$\texttt{replier}_{following\_count}$	-0.058	-0.077	-0.061
$\texttt{replier}_{tweet\_count}$	-0.065	-0.082	-0.028
$\texttt{root\_user}_{followers\_count}$	0.284	0.817	-0.356
$\texttt{root\_user}_{following\_count}$	-0.020	0.032	-0.119
$\texttt{root\_user}_{tweet\_count}$	-0.074	-0.092	-0.019
$\texttt{root}_{reply\_count}$	0.525	1.978	-0.304
$\texttt{root}_{retweet\_count}$	0.276	0.719	0.220
$\texttt{root}_{quote\_count}$	-0.199	-0.378	-0.164
$\texttt{root}_{like\_count}$	-0.292	-1.450	0.788

Table 1: Logistic Regression results for CPIs. Coefficients are underlined if the sign is different in two camps.

RQ1: What are the characteristics of users in CPI?

- **User-Specific CPI Rates** using sparse lasso regression
  - x_i^T is the vector of predictors for the i^{th} observation,
  - β is the vector of coefficients X_i,
  - λ is the regularization parameter to control the degree of shrinkage for the coefficients,
  - y_i is the response variable,
  - $\hat{\beta}$ is the final learned set of coefficients.
- Rows are interactions, columns are users, cells indicate if interaction associates with that user

tweet_id	user_1	user_2	...	user_n	is_CPI
tw1					1
tw2					-1
tw3					1

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{N} \sum_{i=1}^{N} (y_i - \mathbf{x}_i^T \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

RQ1: What are the characteristics of users in CPI?

- **User CPI involvement** strongly correlates (r = 0.78) with their regression coefficients.
- **Democrats** are less likely to engage in CPIs, showing more negative coefficient values.
- **Republicans** are more likely to participate in CPIs based on coefficient distributions.
- **Moderate users** (political center) show the highest CPI rates.
- **Extremists on both sides** are less likely to engage in CPIs.

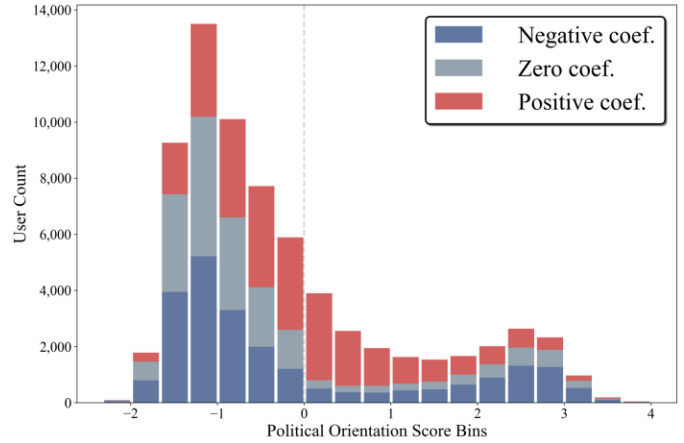


Figure 2: The distribution of political orientation scores for Lasso coefficients of authors on their CPI involvement.

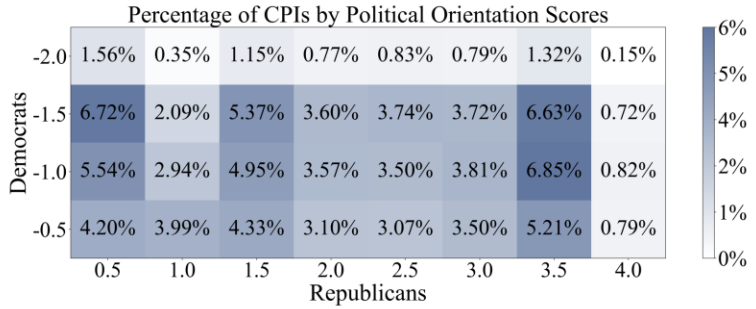


Figure 3: The distribution of CPIs by authors' political orientation scores, grouped into 0.5 interval bins.

RQ1: What are the characteristics of users in CPI?

CPI Toxicity

- **CPI replies are consistently more toxic.**
- **Toxicity decreases as users are closer to the political center.**
- **Center-aligned users show the healthiest dialogue.**
- **Toxicity scores increases with political extremity.**.
- The overall **toxicity curve is symmetric**.
- **Robust to changes in toxicity thresholds (0.4 to 0.9).**

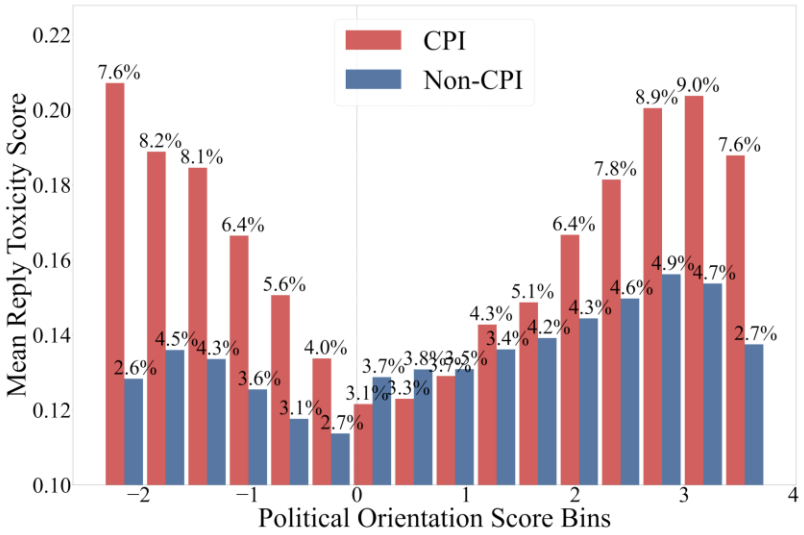
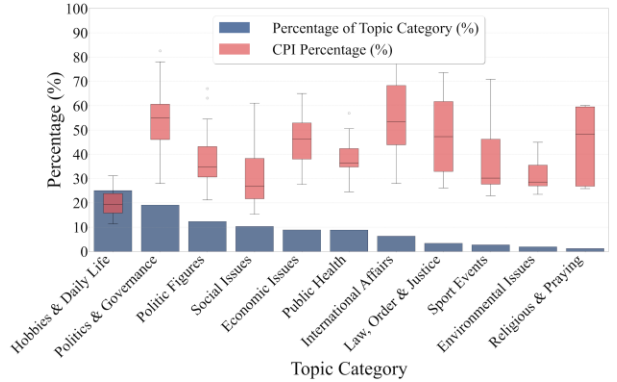


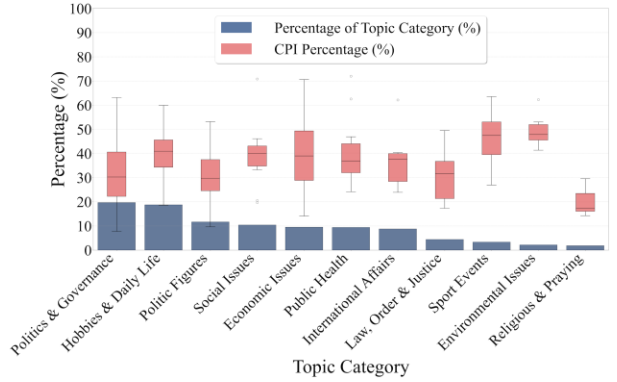
Figure 4: Mean toxicity levels of the repliers' political orientation score bins for both CPI and non-CPIs. The numbers on bars show the percentage of the content that is counted as toxic (score > 0.6) in each bin

RQ2: Which topics are more prevalent in CPI?

- Applied **BERTopic** with BERT embeddings, UMAP, and HDBSCAN to identify 177 distinct topics from root tweets.
- Used **ChatGPT-4o** to generate **human-readable titles** for each topic based on keywords and representative tweets.
- **Manually grouped topics into 11 categories** (e.g., Politics, Sports, Daily Life) to analyze CPI and toxicity patterns across thematic areas.



Democrats



Republicans

Figure 5: The distribution of topic categories in the dataset with the box-and-whisker plot of their CPI percentages on different topics for the entire dataset and both camps.

RQ2: Which topics are more prevalent in CPI?

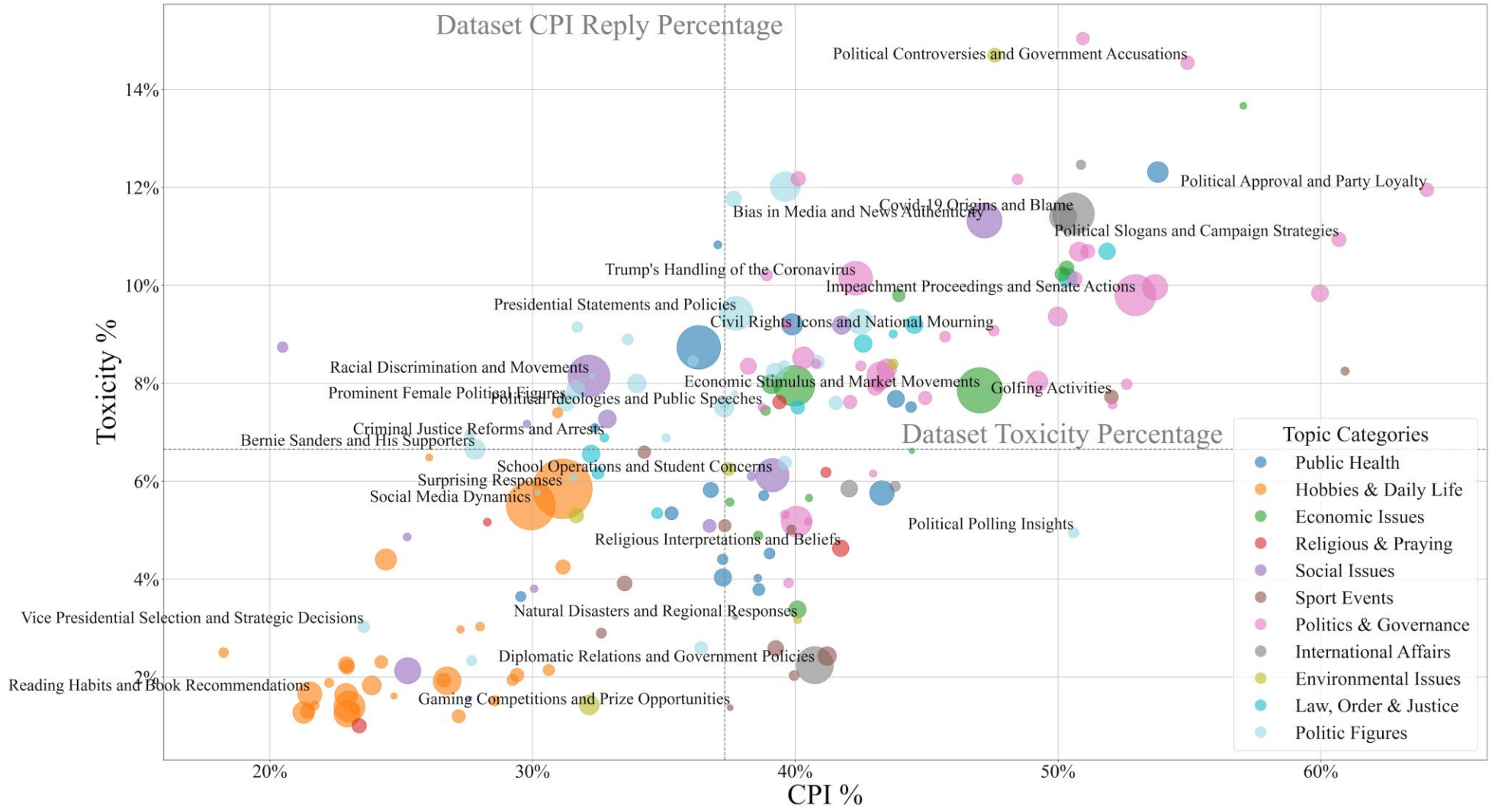
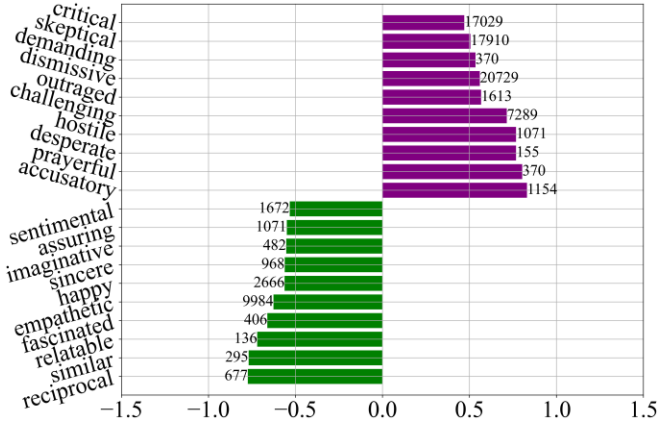


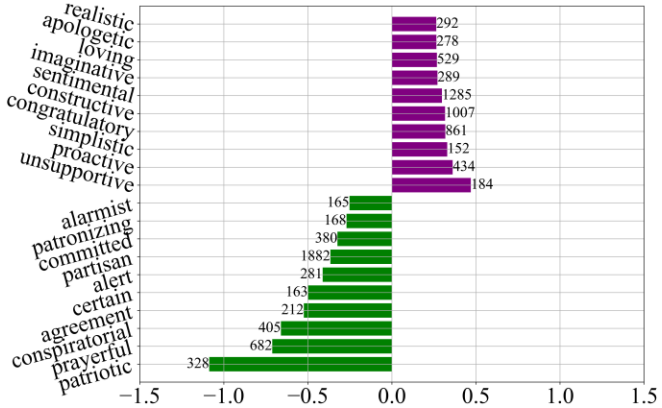
Figure 6: The distribution of the 177 topics with axes representing the CPI percentage and the toxicity.

RQ3: What type of stance appear in CPI?

- **LLM-annotated stance adjectives** are used.
- **Democrats show strong tone shift**: Positive in-group stances turn negative in cross-partisan replies (e.g., “empathetic” → “accusatory”).
- **Democrats’ CPI replies** often use critical and hostile language.
- **Republicans occasionally express warmth** (e.g., “loving”, “congratulatory”) even in CPI replies.



(a) D → D vs. D → R



(b) R → R vs. R → D

Figure 8: Stance-wise differences of partisan vs. crosspartisan replies across parties. The bar labels indicate the overall frequency of the annotation.

RQ3: What type of stance appear in CPI?

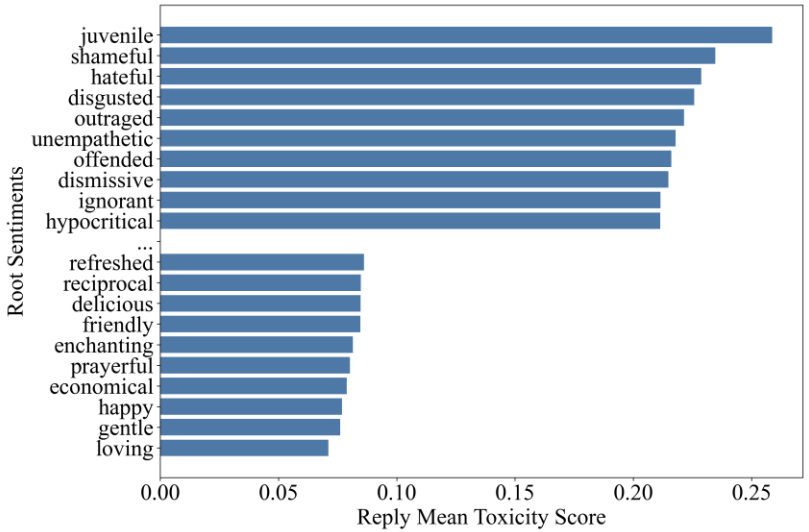
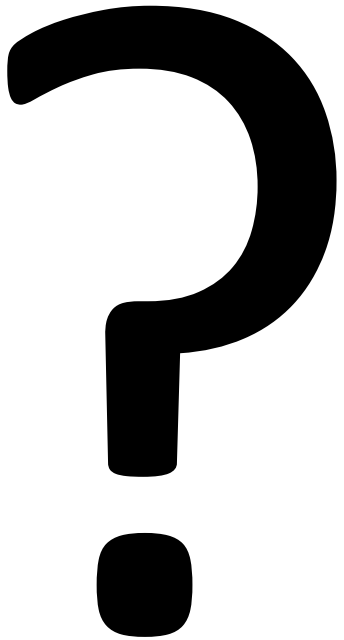


Figure 10: Top 10 root sentiments that attract the highest and lowest toxicity score replies.

Conclusion

- **User popularity and topic sensitivity** jointly shape the likelihood and civility of CPIs.
- **Tone matters:** Tweets with gentle or optimistic sentiment receive less toxic and more cooperative replies.
- **Not all political disagreement is toxic**—certain topics and tones can enable respectful cross-party conversations.
- Understanding CPIs equips us **to detect polarization hotspots and strategically foster bridges** to **counteract manipulative actors** and promote healthier online ecosystems

Questions



Social Media Inference & Manipulation

Dr. Yusuf Mucahit Cetinkaya

Postdoctoral Scholar
The University of Edinburgh

May 29, 2025 - Edinburgh