CommunityLM
Probing Partisan Worldviews from Language Models

Hang Jiang, Doug Beeferman, Brandon Roy, Deb Roy
MIT Center for Constructive Communication, MIT Media Lab

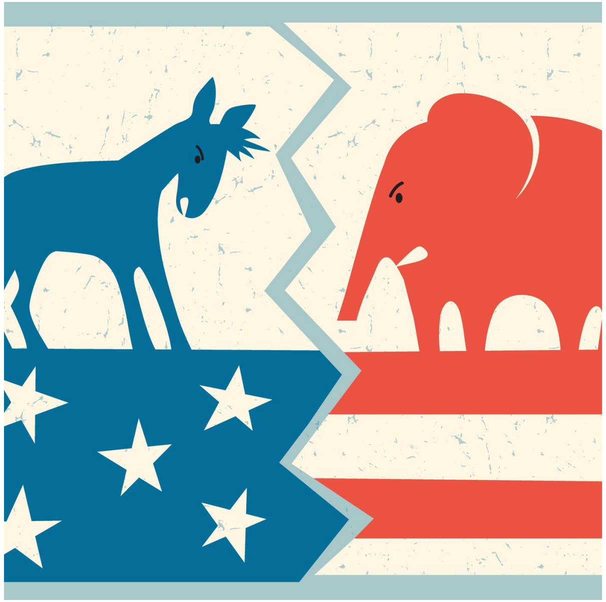
COLING 2022

Motivation

The ever-widening polarization between the US political parties is accelerated by an erosion of mutual understanding between them.

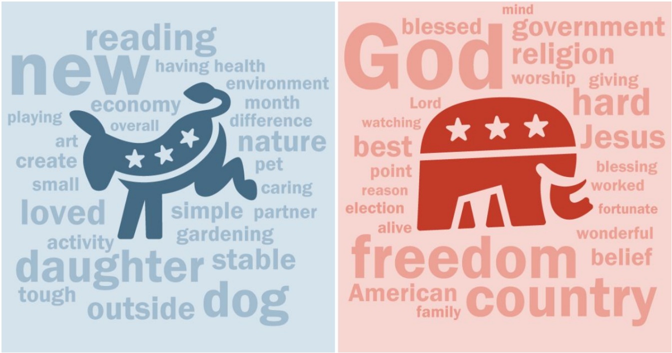
We aim to:

- provide a simple and flexible interface to probe community insights
- encourage constructive dialogue between communities



Previous work on polarized language

1. Li et al. (2017) and R. KhudaBukhsh et al. (2021) use Word2vec to show the left and right use words differently
2. Milbauer et al. (2021) extended the method to 32 communities to uncover ideological differences
3. Palakodety et al. (2020) used a fine-tuned BERT model with fill-in-the-blank cloze statements to mine insights
4. Feldman et al. (2021) fine-tuned GPT-2 on COVID-19 tweet corpora to mine user opinions



However, none of them fine-tune GPT-style language models on community data to probe community worldviews.

Contributions

1. Present CommunityLM based on GPT-2 to mine community insights
2. Evaluate models on ANES to show that models predict community stance
3. Analyze model errors and demonstrate its capability to rank public figures

Check out the GitHub!

Training – Partisan Twitter Data

1. Sample **~1M** active U.S. Twitter users before and after the 2020 presidential election
2. Estimate the party affiliation of Twitter users from the political accounts they follow (Volkova et al., 2014; Demszky et al., 2019)
3. Sample **4.7M tweets (100M words)** from both partisan communities between 2019-01-01 and 2020-04-10

Evaluation – American National Election Studies (ANES)

Please look at the graphic below.

Public Figures

[fttrump1] How would you rate **Donald Trump**?

[ftobama1] How would you rate **Barack Obama**?

[ftbiden1] How would you rate **Joe Biden**?

Social Groups

[ftillegal] How would you rate **illegal immigrants**?

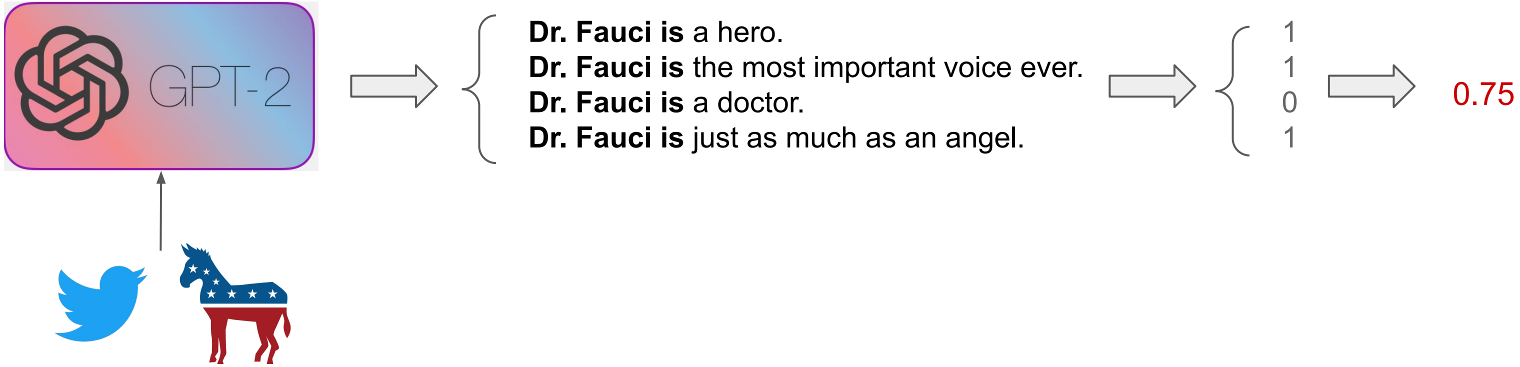
[ftfeminists] How would you rate **feminists**?

[ftmetoo] How would you rate **the #MeToo movement**?

CommunityLM **Framework**

1. Fine-tune GPT language models on community data
2. Design prompts based on survey questions
3. Generate community responses with language models
4. Aggregate community stance based on responses

Prompt	Model	Top 5 Words
Dr. Fauci is a	Republican GPT-2	liar (2.96%), joke (2.67%), hero (2.13%), doctor (1.62%), great (1.61%)
	Democratic GPT-2	hero (10.36%), true (3.63%), national (2.08%), physician (2.06%), great (1.93%)



Baselines

1. Frequency Model
2. Keyword Retrieval (full)
3. Keyword Retrieval (surname)
4. Pre-trained GPT-2 (124M)
5. Pre-trained GPT-3 Curie

Keyword	Question	Dem	Repub
Asian people	ftasian	81	21
Joe Biden	ftbiden1	4177	5377
big business	ftbigbusiness	321	291
Black people	ftblack	3199	1278
Pete Buttigieg	ftbuttigieg1	982	521
capitalists	ftcapitalists	279	197
the Democratic Party	ftdemocraticparty	2094	2646
Anthony Fauci	ftfauci1	102	85
feminists	ftfeminists	351	628
Nikki Haley	fthaley1	169	274
Kamala Harris	ftharris1	1711	1450
Hispanic people	fthisp	28	21
illegal immigrants	ftillegal	251	2233
Amy Klobuchar	ftklobuchar1	451	193
labor unions	ftlaborunions	68	27
the #MeToo movement	ftmetoo	103	84
Barack Obama	ftobama1	684	929
Alexandria Ocasio-Cortez	ftocasioc1	410	534
Nancy Pelosi	ftpelosi1	1467	3549
Mike Pence	ftpence1	911	502
the Republican Party	ftrepublicanparty	1681	838
Marco Rubio	ftrubio1	166	132
Bernie Sanders	ftsanders1	4572	2711
socialists	ftsocialists	627	2697
Clarence Thomas	ftthomas1	157	132
transgender people	fttransppl	165	38
Donald Trump	fttrump1	8501	5479
Elizabeth Warren	ftwarren1	3132	1897
White people	ftwhite	3625	1862
Andrew Yang	ftyang1	585	249

Full name

Keyword	Question	Dem	Repub
Asian	ftasian	2961	1917
Biden	ftbiden1	26558	21748
big business	ftbigbusiness	321	291
Black people	ftblack	3199	1278
Buttigieg	ftbuttigieg1	3514	1348
capitalist	ftcapitalists	1393	941
Democratic Party	ftdemocraticparty	2677	3611
Fauci	ftfauci1	931	1219
feminist	ftfeminists	1686	1470
Haley	fthaley1	531	712
Harris	ftharris1	6753	5416
Hispanic	fthisp	1173	1693
illegal immigrant	ftillegal	312	2815
Klobuchar	ftklobuchar1	1958	584
labor union	ftlaborunions	110	47
#MeToo movement	ftmetoo	114	102
Obama	ftobama1	15390	33105
Ocasio-Cortez	ftocasioc1	751	1792
Pelosi	ftpelosi1	5985	15844
Pence	ftpence1	5818	3021
Republican Party	ftrepublicanparty	2251	1079
Rubio	ftrubio1	508	502
Sanders	ftsanders1	16001	6568
socialist	ftsocialists	3182	12606
Thomas	ftthomas1	2316	3348
transgender	fttransppl	1309	1469
Trump	fttrump1	188170	150589
Warren	ftwarren1	18954	6969
White people	ftwhite	3625	1862
Yang	ftyang1	4443	1433

Surname

Performance on ANES

Model	Prompt	Accuracy	Weighted F1
Frequency Model	—	53.33	54.50
Keyword Retrieval (Full)	—	86.67	87.00
Keyword Retrieval (Surname)	—	93.33	93.33
Pre-trained GPT-2	"[CONTEXT] + X"	74.00±2.79	66.52±5.56
Pre-trained GPT-2	"[CONTEXT] + X is/are"	72.00±1.83	64.63±2.35
Pre-trained GPT-2	"[CONTEXT] + X is/are a"	75.33±1.83	68.47±3.35
Pre-trained GPT-2	"[CONTEXT] + X is/are the"	77.33±2.79	74.71±3.22
Pre-trained GPT-3 Curie	"[CONTEXT] + X"	83.33	83.88
Pre-trained GPT-3 Curie	"[CONTEXT] + X is/are"	93.33	93.50
Pre-trained GPT-3 Curie	"[CONTEXT] + X is/are a"	83.33	83.88
Pre-trained GPT-3 Curie	"[CONTEXT] + X is/are the"	83.33	84.02
Trained COMMUNITYLM	"X"	90.00±0.00	89.63±0.27
Trained COMMUNITYLM	"X is/are"	90.00±0.00	89.82±0.00
Trained COMMUNITYLM	"X is/are a"	86.00±1.49	86.25±1.50
Trained COMMUNITYLM	"X is/are the"	90.67±2.79	90.49±2.68
Fine-tuned COMMUNITYLM	"X"	84.67±2.98	84.46±3.18
Fine-tuned COMMUNITYLM	"X is/are"	96.00±2.79	96.00±2.79
Fine-tuned COMMUNITYLM	"X is/are a"	91.33±1.83	90.83±2.05
Fine-tuned COMMUNITYLM	"X is/are the"	**97.33±1.49**	**97.29±1.52**

Main findings

1. Fine-tuned CommunityLM with "X is/are the" prompt achieves the best performance
2. Fine-tuning >> Training from scratch
3. Fine-tuned GPT-2 >> pre-trained GPT-3 Curie

Error Analysis

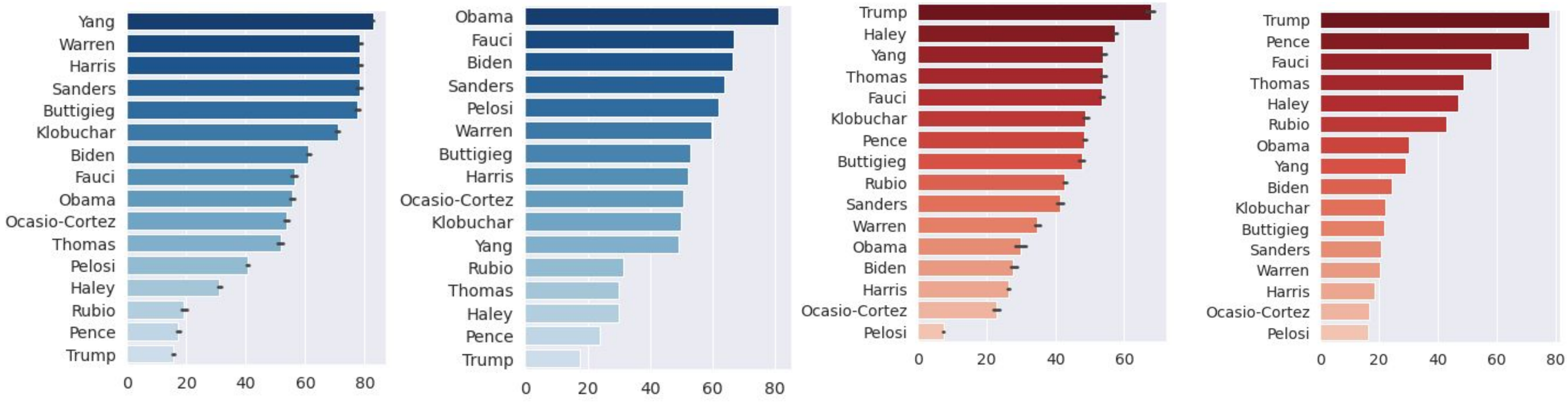
What do the models miss?

1. **Keyword Retrieval (surname)**
    a. “illegal immigrants” and “big business”
2. **Fine-tuned CommunityLM (“X is/are the”)**
    a. “White people”
3. **Pre-trained GPT-3 (“X is/are the”)**
    a. “Dr. Anthony Fauci” and “Asian people”

Top 5 items with the closest average ratings between partisans:

1. Asian people (5.5%)
2. White people (5.9%)
3. Hispanic people (7.7%)
4. Dr. Anthony Fauci (8.4%)
5. Black people (9.7%)

Ranking public figures



CommunityLM Democrat | Gold Democrat | CommunityLM Republican | Gold Republican

Conclusion

1. We present a simple CommunityLM framework to evaluate the viability of fine-tuned GPT-2 community language models in mining community insights.
2. We adopt ANES survey questions and experiment with four types of prompts to generate community responses through GPT-2.
3. We show that generated opinions from CommunityLM are predictive about which community is more favorable towards selected public figures and groups.
4. Our results show that fine-tuned CommunityLM (GPT-2) outperforms the baseline methods.
5. We analyze the model errors and run qualitative analyses to demonstrate that GPT-2 community language models can be used to rank public figures and probe word choices.

Check out the GitHub!

Ethical Concerns

1. The intention of our research is encourage people to **escape from their echo chambers**, hear voices from other communities, and **engage in constructive communication**.
2. We would like to emphasize that our model is **no substitute for deeper engagement with a community**; as discussed in the limitation paragraph, the language model is just an entry point for understanding a community's perspective.
3. Any automated or semi-automated prediction system risks misinterpreting or "erasing" an expressed opinion, and we show in our work that the simpler methods of doing so are more error-prone, and hence measurably more unfair than the proposed approach in the paper.

Limitations and Future Work

- Language models can synthesize unreliable responses
- Language models are shown to be sensitive to prompt design
- We focus on the classic red and blue polarization and do not consider a more fine-grained segmentation of U.S. politics

Acknowledgement

- Belén Carolina Saldías Fuentes, Will Brannon, Suyash Fulay, Wonjune Kang, Hope Schroeder, and Shayne Longpre for their feedback in the early stages of the project.
- MIT Center for Constructive Communication (MIT CCC)

center for constructive communication

mit media lab

COLING 2022