CommunityLM: Probing Partisan Worldviews from Language Models

Hang Jiang, Doug Beeferman, Brandon Roy, Deb Roy | {hjian42, dougb5, bcroy, dkroy} @ media.mit.edu

center for constructive communication | mit media lab

COLING 2022

Introduction

As political attitudes have diverged ideologically in the United States, political speech has diverged linguistically. The ever-widening polarization between the US political parties is accelerated by an erosion of mutual understanding between them. We aim to make these communities more comprehensible to each other with a framework that probes community-specific responses to the same survey questions using community language models (CommunityLM).

Contributions

- We present a simple CommunityLM framework based on GPT-2 to mine community insights by fine-tuning or training the model on community data. This study focuses on Democrat and Republican communities on Twitter but can be easily extended to probe insights from any community based on their public discourse.
- We use ANES questions as prompts and find that GPT-generated opinions are predictive of community stance towards public figures and groups. We experiment with 4 types of prompts and find that the fine-tuned COMMUNITYLM with an "X is the" prompt outperforms all the baselines (including pre-trained GPT-3 Curie) in predicting community stance.
- We analyze the errors made by community language models and demonstrate the capability of the models to rank public figures.

Training – Partisan Twitter Data

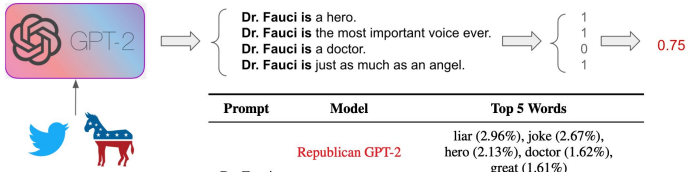
- Sample ~1M active U.S. Twitter users before and after the 2020 presidential election
- Estimate the party affiliation of Twitter users from the political accounts they follow (Volkova et al., 2014; Demszky et al., 2019)
- Sample 4.7M tweets (100M words) from both partisan communities between 2019-01-01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

Evaluation – American National Election Studies (ANES)

- The American National Election Studies (ANES) are academically-run national surveys of voters in the United States.
- We adopt the ANES 2020 Exploratory Testing Survey conducted between April 10, 2020 and April 18, 2020 on 3,080 adult citizens in the US.
- We adapt all 30 questions from "FEELING THERMOMETERS" section of the ANES survey, which asks participants to rate people or groups from 0 ("not favorable") to 100 ("favorable") with the format "How would you rate ____?"

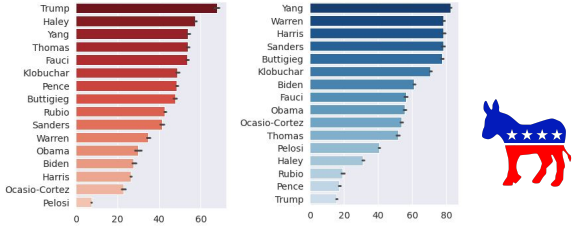
CommunityLM Framework

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%)

Ranking Public Figures



Model 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**

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

Conclusion & Future Work

- We present a simple CommunityLM framework and evaluate the viability of fine-tuned GPT-2 community language models in mining community insights on ANES survey data.
- However, LMs can can synthesize unreliable responses and be sensitive to prompt design. Besides, we only focus on the classic red and blue polarization and do not consider a more fine-grained segmentation of U.S. politics.

References

- Dorottya Demszky, Nikhil Garg, Rob Voigt, James Zou, Jesse Shapiro, Matthew Gentzkow, and Dan Jurafsky. 2019. Analyzing polarization in social media: Method and application to tweets on 21 mass shootings. NAACL 2019.
- Shriphani Palakodety, Ashiqur R KhudaBukhsh, and Jaime G Carbonell. 2020. Mining insights from large-scale corpora using fine-tuned language models. ECAI 2020.